EDITORIALS

DEFAMATION Reforms to libel law unveiled in UK Parliament **p.280**

WORLD VIEW Evolution scientists can do more to counter creationists **p.281**

ASTEROID SCIENCE Giant space rock Vesta is more like a planet than an asteroid **p.282**

Misplaced childhood

The US National Institutes of Health should rethink plans to limit a nationwide study of children. It must not miss a rare opportunity to probe the causes of childhood diseases.

The US National Children's Study is at a crucial turning point. Established by Congress in 2000, the project is a highly ambitious, prospective study of biological and environmental influences on the health of 100,000 US children from before birth to age 21. Twelve years on, the National Institutes of Health (NIH) in Bethesda, Maryland, has spent almost US$1 billion to recruit roughly 4,000 participants (see page 287). Contrast that with the Norwegian Mother and Child Cohort Study, which recruited some 109,000 children — plus 163,000 of their parents — over a decade from 1999 at a cost of about $60 million.

The NIH, rightly, decided that something must change. To continue on this trajectory would be ruinous — both for the agency's finances and for the scientific goals of the project. In response, it proposed dramatic changes earlier this year. These include a plan to recruit children through prenatal care providers, rather than by calling house to house. The NIH also wants to abandon the use of a sample of children designed by statisticians to represent all regions and population subgroups of the country, which would produce findings on both exposures and disease that could be applied to all US children. Instead, it has proposed to draw the bulk of subjects from health-maintenance organizations and large health-care providers.

The first move, away from door-to-door recruitment, makes perfect sense. In a sampling statistician's ideal world, that strategy would be the gold standard. But, as data from the pilot phase of the study show and most of those involved agree, it is simply too expensive. And it is slow. At current rates, it would take so long to recruit children that, given the rapidly changing nature of environmental exposures, it would undermine the scientific value of the study.

The second move, to dispense with a true national probability sample, is more troubling — and much more controversial. The US Institute of Medicine described the sampling method as one of the study's key strengths. The many critics of dropping it now include Edward Sondik, director of the National Center for Health Statistics at the Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia.

It was CDC statisticians who developed the original list of 105 far-flung study locations that the NIH now proposes abandoning. The range of places is important because it avoids selection biases that could lead to invalid inferences, and it represents the broad range of exposures and outcomes in a vast and demographically diverse nation.

The NIH should strengthen its commitment to gathering recruits from these diverse locations. Focusing on health-maintenance organizations, which tend to correspond to large population centres, will introduce inevitable sampling errors. It will also miss a rare opportunity to gather universally useful data on the incidence and prevalence of a plethora of childhood ailments and exposures.

Data on the true burden of many childhood illnesses is inadequate or absent. Yet this information forms the necessary basis to generate hypotheses about environmental exposures. And, as Congress intended, it is needed to discover the factors that contribute to rare and common conditions. By contrast, a skewed sample from large health-care providers, although undoubtedly more convenient and less costly, could produce estimates of disease risk that would not represent the face of the United States. For example, Kaiser Permanente — a large health-maintenance organization based in Oakland, California — might provide ready access to tens of thousands of potential recruits, but none of them lives in South Dakota or Minnesota.

The current NIH managers of the study say that they are closing in on a final sampling strategy for the main study, which is supposed to launch next year.

"The NIH is damaging the public goodwill that is essential to the study's success."

But their inconsistent messages inspire neither confidence nor trust. Although they tell a concerned public that a national probability sample is still under consideration, they told the Senate in a private document in April that the option is off the table. And by appearing to jettison their commitment to run the study in the 105 locations they announced with much fanfare several years ago, they are damaging the public goodwill that is essential to the study's long-term success.

The NIH should state unequivocally — and soon — whether it intends the full survey to be based on the national probability sample. If it concludes that doing so is financially untenable, then Congress should question whether the project will deliver what it asked for, and whether the tens of millions of dollars that flow annually to the study would be better spent on investigator-initiated grants that, however unpredictably, will move health and medicine forward for the coming generations. ■

Needless conflict

Independent experts should be kept from undue suspicion as well as undue influence.

We are what we eat. So it should come as no surprise that food-related issues such as bovine spongiform encephalopathy (BSE), bisphenol A contamination, foot-and-mouth disease, *Escherichia coli* outbreaks and genetic modification resonate with the public. It is unfortunate, then, that discussion of them is often clouded by controversies over the impartiality of scientific advice and whether government regulations are truly unaffected by industry interests.

Questions of food safety, nutrition and agriculture elicit more emotion and public mistrust than almost any other science-based issue. The firestorm over obesity, for example, ignited once again in the United States last week, when the Institute of Medicine issued a

report of nearly 500 pages that makes a compelling case that individual choice is not sufficient to prevent obesity in the current environment of inexpensive high-calorie foods and drinks. The report recommends that industry and government take action to get cheap healthy foods into supermarkets and schools, and that the government intervene to ensure that the right dietary messages get through the flood of advertising. The report, of course, was criticized by the industry forces that would have the most to lose if such changes were implemented.

In this highly charged environment, a controversy over alleged conflicts of interest at the top of the European Food Safety Authority (EFSA) has led to media headlines, criticisms from the European Parliament and a feeding frenzy by some non-governmental organizations critical of EFSA (see page 294). Some of those rushing to judge EFSA might do well to remember, however, that whatever the body's shortcomings, it represents a marked improvement on what went before.

EFSA, which is based in Parma, Italy, was created in 2002 in the wake of the BSE scandal and other food crises. Public confidence in experts and governments had evaporated after it emerged that contaminated beef could cause new variant Creutzfeldt–Jakob disease in humans. At fault was a system in which economic imperatives too often blinkered experts and government ministries — not least departments of agriculture — in their assessment of risks and precautions. EFSA was created to change all that, as an independent agency that would provide scientific advice to the European Union and its member states, entirely separate from those responsible for making decisions. Not even the US Food and Drug Administration enjoys that degree of potential freedom from interference: it uses advisory panels of outside experts, but is ultimately part of a government department. This was made clear last year, when President Barack Obama's administration overruled the agency's decision to make the contraceptive Plan B One-Step (levonorgestrel) available to girls under 17 without a prescription (see *Nature* **480,** 413; 2011).

The powerful agrofood industry will always seek to influence policy, whether within EFSA, or in the European Commission, the European Parliament and national ministries that actually make the decisions. As in other technological industries, many experts have industry links, and scientists' own perceptions of risk can be biased by a pro-technology outlook that might, for example, lead them to be too enthusiastic about certain transgenic crops.

The safeguards against influence and bias should be the same everywhere: comprehensive and timely declaration of potential competing interests, transparency in decision-making, open airing of dissenting opinions and credible independent oversight. EFSA has taken many steps to implement such safeguards, and there seems to be little evidence that it is more affected than any other food-safety body by undue interest.

"Overseers must take care not to unfairly tar the reputations of scientific experts."

The media, non-governmental organizations and elected representatives and their institutions all have important oversight roles. But they also have a responsibility to keep concerns in perspective, and to avoid using them to further personal agendas. Overseers must take care not to unfairly tar the reputations of the many scientific experts who give their time generously and in complete independence to further public-health and science-based decision-making.

The public response to the 2009 swine-flu pandemic points to the risks of unsubstantiated suspicion of scientific advice. There were many wild claims that the medical response to the pandemic was being promoted by industry and industry-influenced experts to sell flu drugs and vaccines. This not only helped to fuel conspiracy theories that the pandemic was a hoax, but also diminished public confidence in health authorities at a time when it was sorely needed.

Advisory bodies must not tolerate shortcomings in procedures to disclose conflicts of interest, but they must defend themselves against any unfair tarnishing of scientific experts. Damage to reputation is extremely dangerous in a society in which the Internet can quickly convert exaggerated claims into supposed facts, and in a political climate in which 'elites' are often suspect. There is more to responsible oversight than just pointing out the problems — real or perceived. ■

Honest opinions

Proposals for a UK law on defamation highlight the power of scientific protest.

Give yourselves a hearty pat on the back. In March last year, *Nature* urged readers who live in the United Kingdom to write to their Member of Parliament with a plea for them to support reforms to the libel laws of England and Wales. Last week, a proposed law that would make most of the sensible and necessary changes was included in the Queen's Speech (an annual to-do list for the British Parliament). With the help of calm seas and a following wind, the libel-law reform, which has broad cross-party support, could be voted in as early as the autumn. (*Nature*'s UK-based readers cannot claim all the credit, of course — the proposed reform comes after a determined and impressive campaign from many individuals and organizations, including the human-rights groups Amnesty International and Global Witness, and the discussion forum Mumsnet.)

Several scientific groups, including the London-based charity Sense about Science, also helped the campaign. Many of the examples that were used to demonstrate that change was needed were scientists who found themselves threatened with legal action for what they viewed as honest academic criticism. *Nature* officially backed the campaign, and, as this issue went to press, still awaits the verdict of a libel suit brought against this journal by Egyptian researcher Mohamed El Naschie.

The proposed legislation directly addresses the concerns of researchers and scientific groups. It would extend a legal defence known as qualified privilege to statements published in peer-reviewed academic journals, as long as they were reviewed by the journal editor and one or more independent experts. This protection would also extend to those who subsequently publish a fair and accurate copy or extract of the original piece.

The new law would also extend the scope of a second existing legal defence against libel — known as fair comment — to cover aspects of scientific practice. Under the proposals, this would help to protect reports of critical statements made at press conferences and academic meetings that are judged to be in the public interest. Those who publish the details of conference proceedings would also be able to draw on this honest opinion defence.

There are other planned changes, too. One is a formal version of a defence currently based on responsible journalism, known in the trade as a Reynolds defence, which helps reporters and publishers to defend a libel claim if they can show, for example, that they checked facts and offered a proper right of reply. And would-be claimants will have to show that their reputation has suffered serious harm. Once in court, however, the burden of proof will remain largely on those who defend libel actions, not on those who prosecute them. Defendants will still have to show that any allegedly defamatory statements are true, which could leave them fighting an uphill battle, albeit equipped with sharper and more numerous weapons.

Still, scientists everywhere should celebrate the planned changes. Journalism on scientific matters has been threatened and stifled for too long. As we wrote in the Editorial in March 2011: "At *Nature*, we have too often been hindered in our core mission because of legal risks." We are not there yet, but we can look forward with optimism. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

WORLD VIEW *A personal take on events*

A. NOVITZKY

Reach out to defend evolution

Creationists seize on any perceived gaps in our knowledge of evolutionary processes. But scientists can and should fight back, says **Russell Garwood**.

Last month, this journal published a fossil study that described a new species of large tyrannosauroid dinosaur covered in feathers. A week later, the US state of Tennessee passed a creationist bill that encourages teachers to discuss the "weaknesses" of evolution. The first event provided fodder for a shrewd and calculated creationist machine; the second was its latest victory. As a palaeontologist, I believe the way that scientists and journals present research in my field can help to feed anti-evolution disinformation. Because we tend to stress novelty and play up scientific disagreement, and like to shift paradigms and break moulds, we offer our critics ammunition. As the events in Tennessee show, the fight against evolution comes with significant consequences. And it goes beyond the United States. The national biology curriculum of Pakistan, for example, dictates that students be taught "that Allah … is the Creator and Sustainer of the universe".

The novelty of a large dinosaur with feathers was a selling point of the recent paper. However, in spite of a widespread agreement on dinosaurs' avian origins, a limited number of researchers remain sceptical. Within days of the paper appearing, the influential creationist organization Answers in Genesis had exploited this disagreement. It misrepresented the *Nature* paper and disagreement about the equivalence of dinosaur feathers and bird feathers, concluding: "Dinosaurs did not evolve into birds … no evidence of feather evolution has been found in the fossil record." It had presented an exciting discovery and a genuine scientific debate (albeit one that has almost run its course) as evidence against evolution, rather than as attempts to refine knowledge in this interesting area.

Another favourite anti-evolution tactic — the god-of-the-gaps — originated in the nineteenth century, and still flies today. This presents perceived gaps in scientific knowledge (genuine or spurious) as evidence in support of theistic world views. The lifeblood of this gappy god is uncertainty — yet good science thrives on unanswered questions. That papers frankly assess and admit shortcomings in current knowledge is vital. But the creationist lobby uses the same literature to try to undermine science.

In my field, uncertainty is everywhere. Much of my work focuses on early land-based animals: the creepy-crawlies that beat our vertebrate ancestors to dry land by a few tens of millions of years. The earliest fossils post-date these first forays into the dry by millions more years. Accordingly, and unsurprisingly, many of them are very well adapted to life on land. Furthermore, although many groups are starkly different from their modern relatives, some look very similar. Take the arachnids called harvestmen. I recently described two fossil examples from rocks 305 million years old that look similar to those we see today.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/stgztf

Their fragility makes harvestmen rare as fossils, and these beautifully preserved new species offered the first opportunity to assess their evolutionary relationships computationally. They turned out to be members of lineages that are still around, and so we reported that harvestmen have an early origin and that they are an example of evolutionary stasis, unlike the majority of other ancient land arthropods, which looked nothing like those we see today.

News coverage of the paper was duly picked up and twisted in the creationist media. A blog on the Lutheran Science website, ignoring the fact that harvestmen are presented as an exception, posited: "And did you note the surprise shown by Dr. Garwood?" It presented stasis as evidence for the non-existence of evolution.

We don't know why harvestmen are such a good example of morphological stasis; but the fact that they are in no way undermines evolution. Rather, it indicates that further work is needed and encourages such work. Yet knowing that unknowns will be presented as evidence of a designer does make writing up the results a potential minefield.

IGNORING THE CREATIONIST THREAT WILL NOT MAKE IT GO AWAY.

We should not let creationist pressure alter the way we do science — the day that researchers become reticent about highlighting inconsistencies and uncertainty would be a dark one. But equally, we are not helpless when it comes to countering creationist disinformation based on our results. I believe that science would benefit greatly if we did more outreach when we publish and publicize our research.

Direct debates with creationists are risky. Organized discussions only support the 'evolution is in crisis' lobby. However, a proliferation of online tools means that we can make accurate information freely available to those interested enough to look for it. Arizona State University's Ask a Biologist web page has fielded more than 25,000 questions from students and teachers since it launched in 1997.

If research is to appear that will attract an obvious creationist interpretation, an accompanying blog post could explain the work and highlight flaws in any anti-evolution attacks. Sites such as the Natural Environment Research Council's Planet Earth Online and the Palaeontological Association-sponsored palaeontologyonline.com provide researchers with vehicles for one-off posts. Publishers can do more, and could offer online summaries in non-technical language, written by the researchers. The open-access journal *Palaeontologia Electronica* already does this.

Ignoring the creationist threat will not make it go away. As scientists, we owe it to the schoolchildren of Tennessee and elsewhere to find another way to beat it. ■

Russell Garwood *is a palaeontologist in the Department of Materials, Harwell Complex, University of Manchester, UK.*
e-mail: russell.garwood@manchester.ac.uk

RESEARCH HIGHLIGHTS

Selections from the scientific literature

CANCER

Environment of chemo success

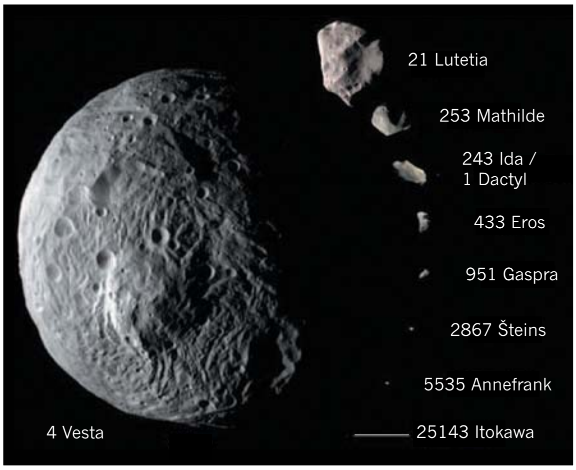
A tumour's response to chemotherapy is shaped by interactions between the tumour and its microenvironment.

Mikala Egeblad at Cold Spring Harbor Laboratory in New York and her colleagues used *in vivo* microscopy to monitor tumours' responses to the chemotherapy drug doxorubicin in mice. They found that the drug is more effective in a mouse model of breast cancer with intermediate tumours rather than precancerous changes or late-stage tumours, and in animals that lack the enzyme MMP9, which acts on the protein matrix surrounding cells and tumours.

The improved drug response seemed to be linked to increased leakage from tumour blood vessels, which facilitates drug access. Moreover, mice lacking a receptor called CCR2 responded more strongly to doxorubicin than did mice with the receptor. Immune cells that express the CCR2 receptor are attracted to tumours as a result of doxorubicin treatment — this can promote tumour regrowth.

Drugs that inhibit MMP9 and CCR2 could be combined with traditional chemotherapies to boost cancer treatment success, the authors suggest.

Cancer Cell **21, 488–503 (2012)**



PLANETARY SCIENCE

Planet-like asteroid

The giant asteroid Vesta resembles a planet more than it does other asteroids, according to Christopher Russell at the University of California, Los Angeles, and his colleagues. In six separate studies, the researchers report their analysis of data from NASA's Dawn spacecraft, which has been orbiting Vesta (pictured, relative to other asteroids) since July 2011.

Vesta was formed about 2 million years after the Solar System's first solid bodies and is the Solar System's second-largest asteroid. The authors report that Vesta is pockmarked with many impact craters, including two overlapping ones, several hundred kilometres wide, at the south pole. One of these polar impacts blasted off material that became more asteroids, known as Vestoids, and meteorites. Shocks from these two big impacts apparently created the troughs that ring Vesta's equator. The asteroid's large size — roughly 260 kilometres in radius — rapid growth and massive iron core may explain how Vesta survived all these collisions.

Science **336, 684–686; 687–690; 690–694; 694–697; 697–700; 700–704 (2012)**

ENVIRONMENTAL HEALTH

Less biodiversity, more allergies

A decrease in the amount of time spent in contact with the natural environment and changes in the population of microbes resident on the skin could be contributing to the increase in inflammatory disorders such as allergies.

To test these ideas, Ilkka Hanski at the University of Helsinki and his colleagues measured immune reactions to common allergens and the composition of skin microbes in 118 adolescents in eastern Finland. The team used the mixture of plants in the young peoples' gardens along with local land use as measures of biodiversity. They found that people living in areas of reduced biodiversity were more prone to allergies, and that allergic individuals had distinct populations of bacteria on their skin. Among healthy individuals, those with a greater abundance of the bacterial genus *Acinetobacter* on the skin produced higher levels of the immunoregulatory protein IL-10, which helps the body to tolerate harmless substances.

Loss of biodiversity could cause problems for public health as well as for the environment, the researchers suggest.

Proc. Natl Acad. Sci. USA **http://dx.doi.org/10.1073/pnas.1205624109 (2012)**

SCIENCE/AAAS

NEUROSCIENCE

Anti-seizure drug boosts memory

One way to improve memory in people with a disorder that can precede Alzheimer's disease is to dampen activity in a part of the brain known as the hippocampus, rather than to boost it as previously thought.

Michela Gallagher at Johns Hopkins University in Baltimore, Maryland, and her colleagues tested the memory of patients with amnestic mild cognitive impairment — in which a person's memory is worse than expected for their age — following treatment with a low dose of an anti-seizure drug, levetiracetam, which dampens excess hippocampal activation. Patients given the drug made fewer mistakes on a memory task and showed less hippocampal activity in brain scans than those given a placebo.

Regulating neural activity could control the progression of Alzheimer's disease, the authors suggest.

Neuron **74, 467–474 (2012)**

BIOMATERIALS

High-voltage plant proteins

Crystals of photosynthetic protein complexes extracted from plant cells can generate extraordinarily high voltages when placed on a conducting surface and stimulated by light.

Each of the light-transducing complexes known as photosystem I

can generate about 1 volt during photosynthesis in the plant. Nathan Nelson and his colleagues at Tel Aviv University measured the electrical potential produced in crystals containing hundreds of layers of photosystem I placed on gold, silicon carbide, or indium tin oxide surfaces. The complexes lined up head to toe, like batteries connected in series. The material produced up to 45 volts when illuminated with laser light, and also generated internal electric fields of up to 100 kilovolts per centimetre — among the strongest ever reported in a crystalline material, even among inorganic semiconductors.

The researchers say that the material could be used to make more efficient high-voltage optoelectronic devices.

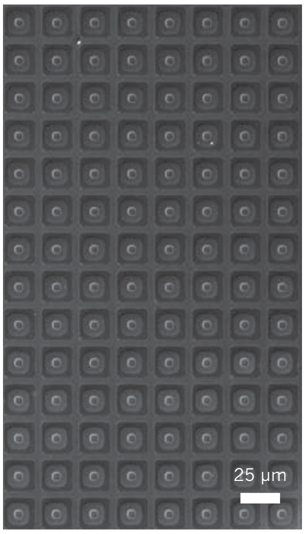
Adv. Mater. **http:/dx.doi.org/10.1002/adma.201200039 (2012)**

PHOTONICS

Solar panel in the eye

Special glasses that fire near-infrared signals onto a device implanted into the retina could one day help to restore vision in blind people. This system would require fewer implanted components such as wires and coils to power the device than other proposed retinal prostheses.

James Loudin at Stanford University in California and his colleagues designed arrays of photovoltaic diodes (**pictured**) that can be inserted into the retina and used to stimulate inner retinal cells. The diodes could ultimately form part of a system whereby a video camera attached to glasses would gather and feed image data to a small portable computer, which would then convert the data into near-infrared light pulses. The glasses would beam those pulses through the eye onto the implanted diodes. This light would deliver both visual information and power to the diodes.



The researchers showed that the photodiode array could activate cells in both healthy and degenerating rat retinas *in vitro* when pulsed with near-infrared light.

Nature Photon. **http://dx.doi.org/10.1038/nphoton.2012.104 (2012)**

ASTRONOMY

Exoplanet signals ring true

Most candidate multi-planet systems spotted by the Kepler space telescope probably contain true exoplanets, according to a statistical analysis.

Kepler spots potential planets beyond our Solar System by looking for tiny dips in brightness as the planets pass in front of their host stars. The method allows the telescope to monitor many stars at once, but can also give false-positive signals.

Assuming that false positives would be randomly distributed among the stars, Jack Lissauer at NASA's Ames Research Center in Moffett Field, California, and his team conducted a statistical test of where candidate planets identified by Kepler are located. They found that more than a third of possible candidates exist as part of multi-planet systems. This is higher than predicted by chance, suggesting that most of these systems contain true planets.

Astrophys. J. 750, **112 (2012)**

COMMUNITY CHOICE

The most viewed papers in science

NEUROSCIENCE

The neural core of consciousness

HIGHLY READ on www.jneurosci.org in April

'Waking up' from an unconscious state requires the activation of only primitive areas deep in the brain — not the higher cortical areas indicated in previous studies on anaesthetized people.

Harry Scheinin at the University of Turku in Finland and his colleagues used position emission tomography to image the brains of 20 volunteers recovering from anaesthesia. The participants had been given either propofol, a standard anaesthetic, or dexmedetomidine, an unusual sedative drug that allows the individual to be awoken temporarily with a shout or a prod. The two-drug study design allowed the researchers to differentiate between brain activities due to the drug and those specifically related to the conscious state.

They found that a core neural network involving just subcortical brain areas and a primitive part of the cortex was activated as subjects recovered sufficient consciousness to respond to verbal instructions.

J. Neurosci. 32, **4935–4943 (2012)**

T. TURNER/NATIONAL GEOGRAPHIC

ANTHROPOLOGY

Ancient Mayan wall calendar

In an underground chamber in Guatemala, archaeologists have discovered the earliest evidence so far of Mayan astronomical tables: dates, numbers and depictions of lunar deities painted or carved on the walls some 1,200 years ago.

William Saturno at Boston University in Massachusetts and his colleagues stumbled across the paintings (**pictured**) while excavating the Mayan city of Xultun. On one wall the researchers found a table containing four columns of numbers, which could represent recurring events related to the cycles of Venus, the Moon, Mars and possibly Mercury. A table on another wall seemed to show 27 columns of dates, each 177 or 178 days apart, with Moon deities at the top of each column. The Maya recorded movements of the Moon in semesters of 177 and 178 days, or six lunar months.

Science 336, **714–717 (2012)**

For a longer story on this research, see go.nature.com/h1pigr

CORRECTION

The story 'Graphene's silicon cousin' (*Nature* **485,** 9; 2012) should have said that silicon was deposited onto a silver surface heated to more than 200 °C. The silicon was heated to more than 1,000 °C.

NATURE.COM

For the latest research published by *Nature* visit:

www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

POLICY

Libel reform

Freedom of speech in academic debate can expect stronger protection in proposed reforms to English libel laws. A defamation bill introduced in Parliament on 10 May gives explicit protection (among other changes) to peer-reviewed statements in scientific or academic journals, and to fair and accurate reports of proceedings at academic conferences. The bill has strong support across political parties, so is likely to pass. The issue was propelled to national attention by campaigners after a 2009 libel case involving the British science writer Simon Singh. See go.nature.com/m1i3ia for more.

Conflict of interest

Diána Bánáti, chairwoman of the European Food Safety Authority's management board, resigned on 8 May to move to the International Life Sciences Institute, a non-governmental organization based in Washington DC and funded by large food and chemical companies. Bánáti has been accused of having a conflict of interest previously, and her move fuelled criticism about the closeness between European science agencies and industrial interest groups. See page 294 and Editorial, page 279, for more.

Global Fund revived

The Global Fund to Fight AIDS, Tuberculosis and Malaria seems to be emerging from a fund-raising crisis, after a thorough restructuring following the resignation of its executive director. General manager Gabriel Jaramillo, appointed in February to overhaul the organization, announced on 9 May that the fund could resume supporting new grants using around US$1.6 billion that will be available up to 2014. The fund, which is based in Geneva, Switzerland, will hand out around $3 billion this year. Five months ago, it had frozen new grants until 2014 after struggling to raise cash from donors, and finding corruption affecting some of its grants.

Nuclear clean-up

Japan's government has announced a ¥1-trillion (US$12.5-billion) plan to bail out the owners of the stricken Fukushima Daiichi nuclear plant, Tokyo Electric Power Company (TEPCO), effectively nationalizing the firm. The deal — inevitable as TEPCO struggled to shoulder the financial burden of cleaning up the plant — will see the government take control of the utility, holding a significant amount of stock and more than half of the voting shares. See go.nature.com/lrbo9f for more.

India drug scandal

India's health ministry has said that it will reform its drug-approvals process and scrutinize clinical trials more closely, after a parliamentary investigation published on 8 May revealed major flaws in the way medicines are tested and cleared for sale. The report said that India's drugs regulator, the Central Drugs Standard Control Organization, was understaffed, lacked expertise, approved drugs that had not been sufficiently tested, and colluded with pharmaceutical companies to speed approvals.

J. LOUTON/SMITHSONIAN INST./NATL MUS. NAT. HIST./EOL.ORG

One million species online

The Encyclopedia of Life, an online database that hopes to create a record of each of the 1.9 million species currently known to biologists, has passed the 1-million species milestone, it announced on 9 May. The project launched in 2008, when it contained just 30,000 species, but over the past few years it has persuaded partners worldwide to upload their databases. The latest jump was provided by data and images from the Smithsonian Institution's National Museum of Natural History in Washington DC; one of the new additions, the dragonfly known as the Sioux Snaketail (*Ophiogomphus smithi*), is pictured. See go.nature.com/2cwvf6 for more.

RESEARCH

Mars rover lives

NASA's Mars rover Opportunity is up and about again, having survived a fifth Martian winter. On 8 May, the rover drove about 3.7 metres downhill from an outcrop called Greeley Haven, on the rim of a massive crater named Endeavour. It had stayed there, low on solar power, since 26 December 2011; in January, it marked its eighth anniversary on the red planet. See go.nature.com/v2olc2 for more.

The end for Envisat

The European Space Agency (ESA) has officially declared its premier environmental satellite dead. ESA unexpectedly lost contact with the €2.3 billion (US$3 billion) Envisat on 8 April (see *Nature* **484**, 423–424; 2012). Images taken by a French satellite showed that Envisat was in a stable orbit, but all attempts to contact it have failed. Mission engineers believe that a critical electrical failure is the most likely cause of its demise, announced on 9 May. Envisat had already operated for more than ten years, double its planned lifetime. See go.nature.com/9ltbg4 for more.

Carbon capture

After years of delays and cost overruns, Norway has opened the world's largest laboratory for testing methods to capture carbon dioxide. The 5.8-billion kroner (US$1-billion) centre in Mongstad, a joint venture between Norwegian energy company Statoil, oil giant Shell and South African petrochemical firm Sasol, was unveiled on 7 May. It can scrub up to 100,000 tonnes of carbon-dioxide emissions from a nearby oil refinery

and a gas-fired power plant, but will vent any carbon dioxide it captures back to the atmosphere. See go.nature.com/wgqg1k for more.

Biodiversity map

An interactive resource for biodiversity analysis was launched on 10 May, promising a new era in visualizing species distributions around the globe. The online Map of Life, funded in part by the US National Science Foundation, will allow users to add or update species data. If the database gains traction in biodiversity circles, it aims to combine mapping with other data — from genomics to environmental variables — to analyse the drivers and impacts of biodiversity shifts over time. See go.nature.com/v8uu74 for more.

PEOPLE

Plagiarism charge

Romania's education and research minister, Ioan Mang, has been accused of plagiarism in at least eight of his academic papers. Mang, a computer scientist at the University of Oradea in Romania, has said that he will resign if experts can prove the allegations, which began circulating on 7 May, shortly after his appointment in the country's new government was announced. See page 289 for more.

D. P. MORRIS/BLOOMBERG/GETTY

Energy head leaves

The branch of the US Department of Energy that specializes in funding high-risk, high-pay-off research is losing its founding director, Arun Majumdar. A mechanical engineer who previously directed the environmental energy department at Lawrence Berkeley National Laboratory in California, Majumdar (**pictured**) has headed the Advanced Research Projects Agency-Energy (ARPA-E) in Washington DC since 2009. But he will leave on 9 June, energy secretary Steven Chu told agency staff on 9 May. Majumdar is moving back to California for family reasons, said agency officials. He will be replaced by Eric Toone, ARPA-E's deputy director of technology.

Physicist convicted

Omid Kokabee, an Iranian doctoral student who has been in jail in Tehran for the past 15 months on suspicion of conspiring against Iran, has been sentenced to 10 years in prison. Kokabee, who was studying laser physics at the University of Texas at Austin, was one of more than ten convicted in a 13 May trial for collaboration with Israel's secret service, Mossad. Close contacts say he was not presented with proof at the trial, and plans to appeal against the sentence. Organizations including the Committee of Concerned Scientists (a human-rights group in New York city) and the American Physical Society (headquartered in College Park, Maryland) had previously asserted Kokabee's innocence. See go.nature.com/jomkwt for more.

Texas resignation

A US$3-billion state-funded cancer institute in Texas is defending the integrity of its grant-making process after its Nobel-prizewinning scientific leader resigned. Alfred Gilman, chief scientific officer of the Austin-based Cancer Prevention & Research Institute of Texas, said on 8 May that he would step down in October. Gilman, who shared a Nobel prize in 1994 for his research on G proteins, said that the funding programme no longer needed him — but cautioned that "negative decisions" at a July funding round could have a "fatal impact" on the institute's peer-review system. See go.nature.com/ntttea for more.

COMING UP

19 MAY
SpaceX of Hawthorne, California, is scheduled to launch its Dragon capsule to take cargo to the International Space Station — a first for a private space-flight firm.
go.nature.com/nosop7

21–23 MAY
Successes and failures in stem-cell therapy are discussed at the World Stem Cells and Regenerative Medicine Congress in London.
go.nature.com/odilmk

22–24 MAY
The future of human space travel is the topic of the Global Space Exploration Conference in Washington DC.
http://glex2012.org/

BUSINESS

HIV prevention

A pill to prevent HIV infection won support from an advisory panel to the US Food and Drug Administration (FDA) on 10 May. Truvada, a combination of the antiretroviral drugs emtricitabine and tenofovir, is currently marketed by Gilead Sciences of Foster City, California, to treat people with HIV. The FDA panel voted in favour of approving it to protect people at risk of contracting the virus, including men who have sex with men, and uninfected people who have HIV-positive partners. A full FDA decision is expected by 15 June. See go.nature.com/mx9xq3 for more.

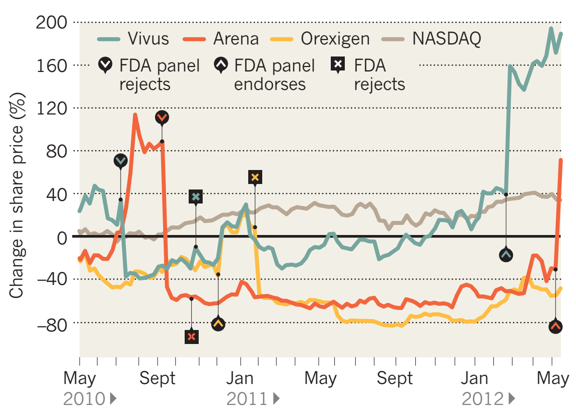
TREND WATCH

SOURCE: NASDAQ

Arena Pharmaceuticals of San Diego, California, saw its stock price almost double on 10 May, when its obesity pill Lorqess (lorcaserin) was recommended for approval by an advisory panel to the US Food and Drug Administration (FDA). In February the panel had backed another diet pill, Qnexa (phentermine plus topiramate), made by Vivus of Mountain View, California. Safety concerns had contributed to rejections of both drugs 18 months ago. Both now await official FDA decisions.

NEW HOPE FOR OBESITY PILLS

Stocks that fell in 2010, when the FDA rejected three companies' weight-loss pills, are surging on new backing from an FDA advisory panel.



NATURE.COM
For daily news updates see:
www.nature.com/news

POLITICS

Plagiarism charge for Romanian minister

Scandal adds to fears that country's research reform is in peril.

BY ALISON ABBOTT

S. MATEI/PHOTOSHOT

Research minister Ioan Mang says that plagiarism allegations against him are politically motivated.

Romania's new government was thrown into turmoil last week after its education and research minister, Ioan Mang, was accused of extensive plagiarism in at least eight of his academic papers.

The allegations first began circulating on 7 May, just hours after Prime Minister Victor Ponta, a Social Democrat, announced the appointment of Mang and other ministers of the new government. Last week, former prime minister Emil Boc, of the Democratic Liberals, called for Mang's resignation, dramatically waving the allegedly plagiarized articles and the original papers in front of television cameras.

The scandal has dismayed many Romanian scientists, who are already nervous that the incoming centre-left coalition government might reverse some of the energizing reforms that were introduced by the previous centre-right coalition to improve the country's sluggish research system.

The radical education and research laws approved last year were designed to introduce competition for positions and research funds, and to eliminate endemic nepotism and other corrupt practices in Romanian academia (see *Nature* **469,** 142–143; 2011). That government also passed a new anti-plagiarism law, which created a Research Ethics Council comprising high-ranking scientists selected by the research minister, and stated that any academic found guilty of such misconduct would automatically lose their job.

Mang is a computer scientist at the University of Oradea in northwestern Romania, and has served on the senate's education committee, which tried to hinder the previous government's research reforms. One of Mang's papers now under scrutiny (I. Mang *Seria Technichni nauki* **12,** 129–135; 2004) is allegedly a near-identical copy of a manuscript intended for presentation at a scientific workshop and authored by cryptographer Eli Biham, the dean of computer science at the Technion Israel Institute of Technology in Haifa.

Biham notes that he had withdrawn the manuscript from the workshop because of conceptual errors in the work, but had been unable to completely remove the document from the Internet. Mang seems "neither to have read nor tried to understand the claims" in the paper, Biham wrote last week on his website.

Mang did not respond to *Nature*'s requests for comment on the allegations, but has told Romanian newspapers that the claims are politically inspired. He has pledged to resign if experts can prove the allegations to be true.

A Romanian blogger, economist Razvan Orasanu at Harvard University's John F. Kennedy School of Government in Cambridge, Massachusetts, contacted authors of some of the allegedly plagiarised papers directly. In addition to Biham, Takeshi Shimoyama at Fujitsu Laboratories in Kawasaki, Japan, and Chu-Hsing Lin at Tunghai University in Taichung City, Taiwan confirmed extensive similarities with their publications.

The Research Ethics Council has taken up the case and is obliged to respond within 90 days. Claiming concern that the independence of the council might be called into question, because its members were chosen by the former education minister, Ponta has also referred the case to the Romanian Academy, even though that body has no mandate to investigate allegations of scientific misconduct.

Former research minister Daniel Funeriu, who designed last year's reforms, says that he is "appalled by the dimension of the fraud", which he says damages the international credibility of Romania at a time when the nation needs to strengthen its research system. In the past 12 months, the previous government issued several calls for competitive research funds worth 1.2 billion lei (US$350 million), under a rigorous system that enlists foreign academics to review project proposals. However, austerity measures have slashed scientists' salaries by up to 25%, and thousands of jobs have been cut in the education and research sectors.

The programme of the new government suggests that it might weaken the research-performance criteria required for academic promotion, for ranking the country's universities and for selecting members of the research ministry's expert committees. "First we must see if they actually do this," cautions Bogdan Dumitrescu, a computer scientist working at the Tampere University of Technology in Finland, and a member of the Romanian national council for attesting titles, diplomas and certificates, which advises on academic promotion. "But this seems to be the way the wind is blowing, and it would be a great pity."

As the investigation into Mang's publications gathers pace, Funeriu argues that "the actual fight is between ethically sound scientists and the demons of the past", referring to the communist era under the dictatorship of Nicolae Ceaușescu. "The world should know that Romanian researchers are not like their current science minister." ■

MORE ONLINE

A. OMELCHENKO/SHUTTERSTOCK

TOP STORY

Restoring sight with wireless implants go.nature.com/hehivm

MORE NEWS

- Some fish stocks certified as 'sustainable' are overfished go.nature.com/9o5bec
- A genetic link to post-traumatic stress go.nature.com/nrgz6a
- Chinese university wins degree of freedom go.nature.com/osyido

WATCH THE MOVIE

Paralysed woman controls robot arm with her mind go.nature.com/qg1c4z

NEWS IN FOCUS

ASTRONOMY γ-ray bursts shine a light on the early Universe **p.290**

MALARIA Gains are threatened by insecticide-resistant mosquitoes **p.293**

EUROPE Are regulators too cozy with industry groups? **p.294**

ECOLOGY How the coyote became North America's top dog **p.296**

ERIN NIEMANN/ELP-STUDIO.COM

Atlas Thiex of South Dakota is one of almost 3,000 children enrolled in the US National Children's Study.

HEALTH RESEARCH

Child-study turmoil leaves bitter taste

Frustration mounts as ambitious US project is scaled back.

BY MEREDITH WADMAN

Trisha Massmann got a jolt when she received a letter informing her of imminent changes to the US National Children's Study (NCS). She had enthusiastically signed up to the project as an expectant mother in the summer of 2009, after one of its recruiters knocked on the door of her blue clapboard house in the farming community of Granite Falls, Minnesota, population 2,881. By the time Massmann's son, Brett, was born the following February, two fieldworkers from the NCS — a hugely ambitious effort to track environmental and biological influences on the health of 100,000 US children from before birth to age 21 — had spent hours in her home collecting, among other things, dust, air, water and toenail clippings from the parents to be. The same researchers would continue to visit and monitor Brett at regular intervals, becoming a fixture in the family's life.

But in the background, the study has been wracked with budget and management problems and has become a headache for its overseers at the US National Institutes of Health (NIH) in Bethesda, Maryland. In March this year, the disharmony rippled out to Granite Falls, where Massmann's letter informed her that starting in July, the study's activities, including family contacts, would be taken over "for an undetermined period of time" by a research-consulting firm based in North Carolina.

Massmann immediately sent a text message to Kari Loft, one of the NCS fieldworkers she had grown to know and consider a friend, saying, "I'm not cool with this." Recalling the episode during a visit from Loft this month, Massmann said that before joining the study, she had received assurances that the same fieldworkers would be with her long-term. "That was really important to me. Because they have built up a relationship with my kid." Now, she says, "I don't know if I even want to do it".

Thousands of parents are facing the same uneasy transition as the NCS grapples with its budget woes and undertakes a wholesale restructuring. The affair has unleashed acrimony at all levels, starting with the first seven pilot sites, or Vanguard Study Centers, including the one at South Dakota State University in Brookings, which recruited Massmann. They are slated to shut down this summer, and scores of fieldworkers, including Loft, will lose their jobs. A further 33 pilot sites will face a similar fate when their contracts with the NIH expire over the next 16 months. The NIH ▶

says that investigators will be eligible to compete for new contracts, but some are sceptical. "The fact is, we're shutting down. We're done," says Jennifer Culhane, principal investigator for the NCS centre at the Children's Hospital of Pennsylvania in Philadelphia, another of the original seven pilot sites.

Separately, turmoil has rocked the study's advisory committee: two members resigned in March, saying that they weren't consulted about changes in sampling strategy that they feel will undermine the study's scientific value. The affair underscores the organizational challenge inherent in running such a large longitudinal study, but more than that, say critics, it also casts doubt on the NIH's commitment to the study's vision, and to its eventual success.

FROM THE CITIES TO THE SWAMPS

The NCS was born in 2000, when Congress directed the NIH to study "the effects of both chronic and intermittent exposures on child health and human development". Law-makers specified that the exposures should be biological, chemical, physical and psychosocial, and that the study should address health disparities and monitor US children in all their diversity.

Statisticians at the Centers for Disease Control and Prevention in Atlanta, Georgia, picked 105 sites (see 'A representative sample') for the study, with each meant to recruit about 1,000 pregnancies. The sites run the gamut from urban California to the swamps of Florida: a diverse sample meant to produce findings that could be generalized to all US children, with 40 Vanguard centres launching first and the main study to follow in 2013. There was plenty for Congress to like in a project with such broad representation, which promised to illuminate the roles of environmental factors in diseases such as asthma, autism and diabetes. Repeated attempts by US president George W. Bush to cut the NCS were always thwarted.

However, by 2009, just as fieldworkers were beginning to enrol subjects for the pilot phase, the Senate was accusing the study's then-leaders of a "breach of trust", saying that they had knowingly underestimated its costs by as much as half. The study's director was ousted and replaced with Steven Hirschfeld, a paediatric oncologist and associate director for clinical research at the National Institute of Child Health and Human Development (NICHD) in Bethesda. As of late February this year, the pilot phase had enrolled 2,850 babies; a further 1,200 are expected by the end of the year. By then, the NCS will have spent US$992 million, or about $250,000 per child.

Many principal investigators at Vanguard sites contend that tens of millions of dollars have been wasted on, for instance, an overly complex data-transmission system that requires huge amounts of local programming time.

But with so much invested in local infrastructure, say the study's defenders, jettisoning the Vanguard centres will only multiply the loss.

The move "is so shortsighted", says Rick Holm, a physician who, as chief of staff at the Brookings Health System, helped to build community support for the study. "We made a huge effort to get this started. And the power of community is local."

That sentiment is echoed by Emily Thiex, whose 11-month-old son, Atlas, is a study participant. "It's a National Children's Study," she says. "They should be in every state." Thiex's family works a cattle and sheep farm outside Brookings, so she is well aware that "there are different environmental factors everywhere".

NARROWED FOCUS

But the NIH has come to see the study as unsustainable. Last month, at a private Senate briefing, Hirschfeld delivered a document stating that the agency will abandon the NCS's goal of representing the United States in a statistically generalizable way. That approach, the document said, could not yield enough subjects "within either a scientifically sound timeframe or a fiscally sound budget". (Speed of enrolment is important because environmental exposures change over time.) Instead, the document said, the main study will use health maintenance organizations (HMOs) and "other health care provider networks" as the primary recruiters. When it came to light, critics complained that the strategy would bias the study: only a distinct subset of the US population is covered by HMOs, and there are none in South Dakota, for example.

"'It's very sad that rural America is getting the short end of this," says Bonny Specker, the principal investigator at the Brookings centre, which has enrolled more than 400 babies so far — more than any other study site. Others share her concern: at a public meeting of NCS advisers on 24 April, 31 Vanguard principal investigators united to present a white paper arguing adamantly for the scientific need for a generalizable sample.

Hirschfeld will meet government statisticians on 29 May before finalizing the sampling method, in consultation with NIH director Francis Collins and Alan Guttmacher, director of the NICHD. Last week, Guttmacher told *Nature*: "We are currently exploring whether a hybrid model that is primarily provider-based could be designed so that it still provides a probability sample that would allow generalizability."

The inconsistency of the messages has troubled Congress. Senator Tim Johnson (Democrat, South Dakota) told *Nature* that he is "disappointed" with the sidelining of the Vanguard centres, including the one in Brookings. He vowed "to ensure the integrity and intent of the study is not compromised".

Senator Thomas Harkin (Democrat, Iowa), chairman of the spending subcommittee that funds the NIH, says: "I am concerned that Congress has appropriated a total of nearly $1 billion for this project and we still do not understand exactly how the NIH plans to implement it. We need some clarity."

Others say that the changes to the study are necessary after years of poor leadership and overgrown ambitions. The NIH "allowed the study to take on a form that was completely untenable", says David Savitz, an epidemiologist at Brown University in Providence, Rhode Island, and former head of a Vanguard centre in North Carolina. "For this thing to work there was going to have to be some sort of upheaval."

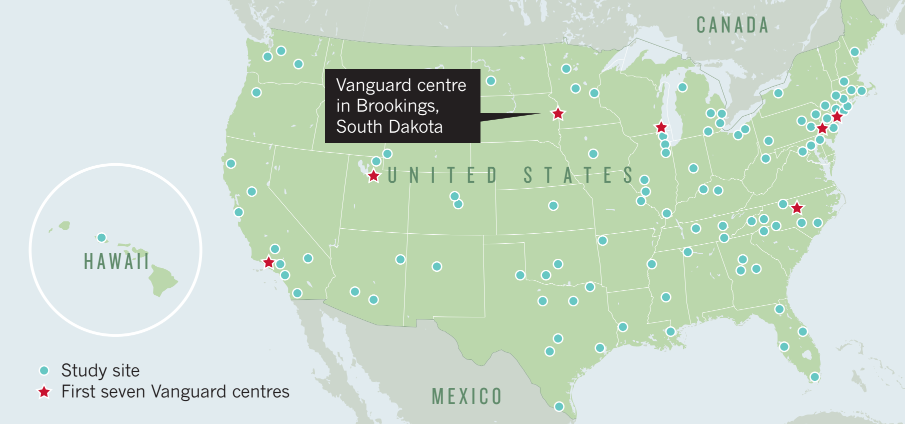
Hirschfeld defends the changes, made against the backdrop of a proposed 15% budget cut for the study in 2013. "Unless we make the adjustments needed to ensure that the study can be carried out successfully, we will not be able to realize the vast potential it has to offer," he says.

That is of little comfort to fieldworkers such as Loft, who was going to a job interview after visiting Trisha Massmann this month. In her car after the visit, she burst into tears.

"I feel like I have lied to her," she says. "I haven't. But the study has let her down. I told her I'd follow her and Brett for 21 years. That won't happen." ■ SEE EDITORIAL P.279

A REPRESENTATIVE SAMPLE

The main US National Children's Study hopes to follow the health of 100,000 children until adulthood, starting in 2013. The 105 original study sites were carefully selected to reflect a range of communities and environmental influences.



ESO

As their beams of intense light shine through surrounding gas, γ-ray bursts like the one illustrated above pick up clues to the Universe's chemical evolution.

ASTROPHYSICS

Messages from the early Universe

Bright and brief, γ-ray bursts hold clues to cosmic history.

BY ERIC HAND

Unimaginably distant and powerful, the brief flashes of high-energy radiation known as γ-ray bursts (GRBs) were once one of astronomy's deepest mysteries. Now they are becoming a penetrating new tool. With orbiting observatories such as NASA's Fermi and Swift spacecraft routinely spotting the bursts, astronomers are laying plans to use them as cosmic flashbulbs to scrutinize the obscure details of the Universe's early years.

Seen almost daily, from all directions in space, GRBs are now thought to signal the collapse of a massive star's core into a black hole, an event that triggers a cataclysmic explosion. Their intense light can shine all the way across the visible Universe — bearing witness to the earliest chapters of its roughly 13-billion-year history. Theorists' understanding of the flashes is still evolving (see 'Flash of insight'), but at the 2012 Fermi/Swift GRB conference last week in Munich, Germany, astronomers discussed how they could use GRBs to chart the chemical evolution of the cosmos as light from the bursts is filtered through gas in the galaxies in which they reside.

Volker Bromm, an astronomer at the University of Texas at Austin, says that GRBs are "cosmic Rosetta stones" that might even carry information about the composition of the Universe's very first stars, a few hundred million years after the Big Bang. "It almost has a metaphysical appeal," he says. "We want to go to the moment of first light."

Along with faint galaxies and quasars — the luminous cores of young galaxies with supermassive black holes at their centres — the objects that emit GRBs are among the most distant in the cosmos. As messengers from the early Universe, GRBs have advantages over the other two, says Nial Tanvir, an astronomer at the University of Leicester, UK. They are much brighter than distant galaxies, which means that a spectrograph has more information to work with when it splits a GRB's glow into its constituent wavelengths to reveal chemical absorption lines. And although quasars shine brightly, their light can be more erratic than that of GRBs and their spectra more complicated, which makes it more difficult to extract information about the material they have shone through.

The challenge is that GRBs are unpredictable and brief — typically lasting only seconds at the highest energies. Their ephemeral flashes are followed by lingering afterglows that can be measured at longer wavelengths, but ground-based observatories must react quickly if they are to pick up the afterglows as soon as a spacecraft detects a burst. However, it can be done: one burst, detected by Swift in September 2005, was so bright that the 8-metre Subaru telescope on Hawaii detected the afterglow and obtained a spectrum more than three days later. With a measured redshift of 6.3, the burst is estimated to have occurred when the Universe was less than 7% its current age. The spectrum, rich in detail, revealed that the re-ionization of hydrogen gas — a key turning point in cosmic history after the Universe cooled and darkened following the Big Bang — had been essentially complete.

➲ NATURE.COM
For more from the early Universe, see:
go.nature.com/azbtmo

But astronomers want to go even further back. GRBs have been going off ever since the formation of the Universe's first stars, which were probably massive, bright and short-lived. The light of such stars when they expire violently as GRBs would offer a coveted chemical fingerprint of the surrounding gas — the primordial stuff of the very early Universe.

By analysing GRBs in galaxies from different epochs, astronomers might be able to trace how the composition of the early Universe evolved, as early generations of stars burned a primordial supply of hydrogen and helium, converting it into heavier elements, collectively termed metals. "When did these big stars start making all these metals? When did they turn on?" asks Neil Gehrels, an astronomer at the Goddard Space Flight Center in Greenbelt, Maryland, and principal investigator for Swift.

To help to get a jump on early GRB observations, Jochen Greiner, an astronomer at the Max Planck Institute for Extraterrestrial Physics in Garching, Germany, and his team built

FLASH OF INSIGHT

What makes γ-ray bursts shine?

Last week, astronomers at a conference in Munich, Germany, presented a new picture of the internal mechanics of γ-ray bursts (GRBs), the fleeting but exceptionally luminous cones of light that jet outward along the rotation axes of stars as they collapse into black holes and explode as supernovae.

With only a flash to go on, the dissection of a GRB's fiery glow is a challenge to astronomers. For decades, theorists were convinced that most of a burst's γ-rays originated with the shock waves that rush outward from the blast at nearly the speed of light. The twisted and compressed magnetic fields embedded within the shocks can accelerate electrons and cause them to emit γ-rays as synchrotron radiation.

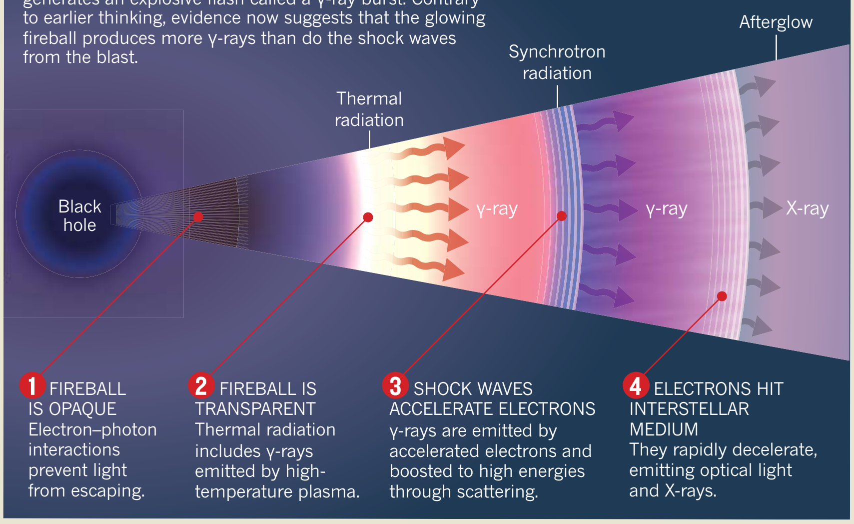
But, as data accumulate, the evidence suggests that most of the γ-rays are emitted as thermal radiation at the scorchingly hot surface, or photosphere, of the exploding fireball (see 'Anatomy of a burst'). In such a scenario, GRBs shine in the same way as stars, through the collective motion of their constituent particles, but at energies that correspond to a temperature in the billions of degrees (see *Nature* http://doi.org/hwv; 2012).

"The rise of the photospheric model for me is transformative," says Julie McEnery, the project scientist for the Fermi γ-ray space telescope at the Goddard Space Flight Center in Greenbelt, Maryland. "It's really a sea change."

Fermi, launched in 2008, has been instrumental in guiding this shift. It does not spot GRBs as precisely as the earlier satellite Swift, but it analyses the shape of a GRB spectrum across most of its γ-ray energies. And the shape is not consistent with synchrotron radiation, says Sylvain Guiriec, a Fermi team member at Goddard. He says he has detected about a dozen spectra from bright GRBs that contain a small bump, a sign that thermal emissions are making a large contribution. **E.H.**

ANATOMY OF A BURST

When a black hole forms from a collapsed stellar core, it generates an explosive flash called a γ-ray burst. Contrary to earlier thinking, evidence now suggests that the glowing fireball produces more γ-rays than do the shock waves from the blast.



the Gamma-Ray Burst Optical/Near-infrared Detector (GROND) and added it to a 2.2-metre telescope operated by the European Southern Observatory (ESO) at La Silla in Chile. GROND responds to alerts from Swift, and seizes control of the ESO telescope. The automated system can make a quick estimate of a burst's distance; if the candidate is remote, Greiner and his colleagues call astronomers at the ESO's nearby Very Large Telescope, which has instruments that can make fine spectroscopic measurements. But Greiner is sometimes unable to convince them to interrupt their work. "They don't realize we have to react in minutes," he says.

Greiner also worries about the fact that Swift, although still working well, was designed to last only two years. Gehrels, however, is optimistic that with more spectrographs on ground-based telescopes, astronomers will be able to make the most of what Swift finds. He believes that it is just a matter of time before an explosion is detected that takes us even closer to the Big Bang. "All it's going to take is one burst," he says. "We just haven't got lucky yet." ■

GLOBAL HEALTH

Malaria surge feared

The WHO releases action plan to tackle the spread of insecticide-resistant mosquitoes.

BY AMY MAXMEN

The war to bring malaria to heel has made slow but steady progress during the past decade, with the overall mortality rate dropping by more than 25% since 2000. A key factor in this progress has been improved control of mosquitoes, which transmit the *Plasmodium* parasite — a potent killer that claimed an estimated 655,000 lives in 2010 alone. But health officials fear that the spread of insecticide-resistant mosquitoes could bring about a resurgence of the disease. To help combat this threat, on 15 May the World Health Organization (WHO), based in Geneva, Switzerland, issued a strategic plan to curb the spread of resistance.

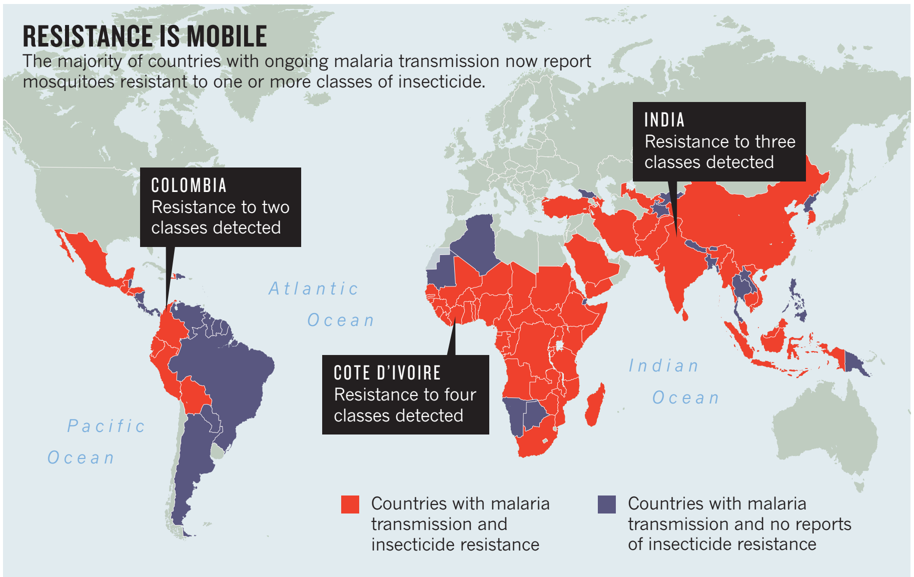
"We don't want to wait for failures to happen," says David Brandling-Bennett, the senior adviser for infectious diseases at the Bill & Melinda Gates Foundation in Seattle, Washington, who advised on the document.

Such failures could reverse the recent drop in malaria mortality credited to insecticide spraying in the home and coating of bed nets, which save about 220,000 children's lives each year, according to the WHO. Insecticide resistance could also result in as many as 26 million further cases a year, the organization predicts, costing an extra US$30 million to $60 million annually for tests and medicines.

The WHO report says that insecticide-resistant mosquitoes already inhabit 64 malaria-ridden countries (see map). The problem is particularly acute in sub-Saharan African countries such as Benin, Burkina Faso, Cameroon, Côte d'Ivoire, Ghana, Ethiopia and Uganda, where mosquitoes are frequently resistant to compounds known as pyrethroids and even to the organochloride DDT, venerable tools of mosquito control. Because they are extremely safe for children, effective against mosquitoes and affordable, pyrethroids are the only insecticides used to treat bed nets, as well as the first choice for household spraying.

> ***"In 2004, there were pockets of resistance in Africa, and now there are pockets of susceptibility."***

Health authorities in Somalia, Sudan and Turkey have also reported sporadic resistance to the two other classes of insecticides recommended by the WHO for safe and effective household spraying: carbamates and organophosphates. Resistance has probably evolved several times independently, and is now spreading as extensive use of pyrethroids and other insecticides favours resistant mosquitoes. "In 2004, there were pockets of resistance in Africa, and now there are pockets of susceptibility," says Janet Hemingway, chief executive of the Innovative Vector Control Consortium (IVCC), a product-development partnership based in the United Kingdom.



Among other things, the WHO recommends rotating the classes of pesticides used to spray houses, and developing safe and effective non-pyrethroid insecticides that can be used to treat bed nets. To implement all of the WHO's suggestions would cost $200 million — on top of the $6 billion that the WHO requested last year to fund existing malaria-control programmes. Rob Newman, director of the Global Malaria Programme at the WHO, hopes that the report will draw more funds to the table as donors grasp the situation. "If we can stop pyrethroid resistance from spreading, it will be cheaper in the long run," Newman says.

But the two largest players in malaria aid — the Global Fund to Fight AIDS, Tuberculosis and Malaria, and the US President's Malaria Initiative (PMI) — have not yet pledged additional money to fight resistance. Their spending on mosquito control is already high — in 2009, 39% of the Global Fund's malaria expenditures went towards insecticide-treated bed nets and household spraying, as did 59% of the PMI's in 2010.

➲ NATURE.COM
To read an Outlook on malaria, see:
go.nature.com/gzyguu

For now, pyrethroids are the only class of insecticides approved by the WHO for bed nets, and where spraying is concerned they are less costly than the alternatives. Vestergaard Frandsen, a company based in Lausanne, Switzerland, says that it has in the pipeline a bed net coated with a non-pyrethroid insecticide — one that does not belong to any of the four WHO-approved classes — and that the company expects to bring this to market within the next five years. It is also one of several companies partnering with the IVCC to create innovative mosquito-control products.

In the meantime, health officials may be able to keep malaria at bay by swapping insecticides. The report notes that in Colombia, for instance, mosquitoes regained susceptibility to pyrethroids after five years of treatment with an organophosphate. But some African countries lack the surveillance needed to spur such an approach. To address that deficiency, the report urges that a global database be set up to track the spread of resistance, and that entomologists be trained and hired at surveillance stations. That could prove the most challenging goal of all. "Nobody wants to fund capacity building," says Newman. "Donors would rather say they purchased $10,000 in bed nets than pay a salary."

African ministers of health realize the need to manage resistance but can't do much without outside funds, explains Maureen Coetzee, a medical entomologist at the University of the Witwatersrand in Johannesburg, South Africa. "In some countries, malaria control means one person sitting in one room, and he's lucky if he's got a chair," she says. ■

POLICY

EU agencies accused of conflicts of interest

European Parliament reprimands food advisory body for industry links.

BY DECLAN BUTLER

Three European agencies are fighting to rebut charges that they enjoy an overly cosy relationship with companies and interest groups.

The latest twist in the saga came last week, as the European Food Safety Authority (EFSA), based in Parma, Italy, was urged by members of the European Parliament to tighten safeguards against potential conflicts of interest among its staff and advisers. In recent months, two other agencies — with responsibility for the environment and for the safety of human and animal medicines — have had to deal with conflict-of-interest allegations that have sparked concerns among some parliament members.

The timing of the row could not have been worse for EFSA, which has just begun rolling out a series of reforms intended to reinforce the independence of the food-safety and nutritional advice that it gives to policy-makers.

NEGATIVE PERCEPTION

On 9 May, EFSA announced that Diána Bánáti, director general of Hungary's Central Food Research Institute in Budapest, had resigned the previous day as chair of the authority's management board, an unpaid position. She had drawn criticism by accepting a full-time job as executive and scientific director of the European branch of the International Life Sciences Institute (ILSI). The institute is a non-governmental organization based in Washington DC that coordinates and pays for research and risk assessments on topics such as food safety and nutrition, and which is funded by large food, chemical and pharmaceutical companies.

The new post was "incompatible" with a code of conduct for board members adopted in June 2011, says Catherine Geslain-Lanéelle, EFSA's executive director. The code stipulates that members must not act in any way that could create a potential conflict of interest or the public perception of one, or harm public trust in the authority. Geslain-Lanéelle complains that Bánáti only informed the authority about her appointment on the day that she signed the contract for her new position at ISLI. "EFSA regrets that this has happened, and the way it happened," she says, adding that the situation risks creating a "negative perception" of the authority.

EFSA

Diána Bánáti denies any conflict of interest in her move to an industry-funded interest group.

Bánáti argues, however, that "this is the usual and accepted way in which people move from one job to another". If EFSA required such notice at an earlier stage, she adds, it "would potentially infringe upon people's ability to manage their own careers".

Following on from controversy over an earlier alleged conflict of interest (*Nature* **467**, 647; 2010), Bánáti had resigned as a member of ILSI's European board of directors in October 2010, and was re-elected as chair of EFSA's board the same month. Since then, she says, "I have met many scientists who work with ILSI as a result of my normal scientific work", but she adds that she had no formal relationship with the institute until it contacted her about the directorship in March. She insists that she has "continued to act in complete accordance not just with EFSA's rules, but my own personal moral code, which means I made decisions and offered opinions completely and solely on the basis of good science".

In March, EFSA had unveiled new rules governing conflicts of interest for its in-house staff, as well as its outside experts, including specified lists of activities that would preclude scientific experts from serving on advisory panels. Scientists previously employed by industry must now have a two-year 'cooling-off' period before they can sit on EFSA's scientific panels, for example, and scientists who receive more than 25% of their research funding from industry face other restrictions on the roles they can undertake at the authority. Former staff — but not scientific advisers — must notify EFSA of all new employment for two years after their departure, and can be asked to refrain from working with the authority in their new job for one year.

In the past few months, EFSA has begun to randomly screen the declarations of interests that its scientists must complete, and it has created a Committee on Conflicts of Interests to investigate any complaints about undue influence. Sue Davies, a vice-chair of EFSA's management board, who is also chief policy adviser at Which?, a UK consumer watchdog, says that the authority's efforts have helped to clarify how potential conflicts of interest should be managed. These can be particularly frequent in an area where industry and regulators often seek out the same experts for guidance.

Davies also emphasizes that EFSA's internal structure prevents its management board from influencing its scientific work. The board's tasks are administrative and strategic, she explains, and although it does oversee the authority's process for appointing external scientific experts to panels, it is not involved in actually choosing them. Board members also have no role in EFSA's scientific deliberations, she adds.

> ***"EFSA's close links to the food lobby undermine the authority's ability to act in the public interest."***

Critics, however, remain unconvinced. In a statement, Nina Holland, a spokeswoman for the Brussels-based Corporate Europe Observatory, a non-governmental organization that campaigns against industry influence on European Union policy, described Bánáti's move as "an absolute scandal". "EFSA's close links to the food lobby through ILSI Europe undermine the authority's ability to act in the public interest," she said.

Members of the European Parliament have also expressed their displeasure, last week voting by a narrow majority to defer approval of EFSA's 2010 budget report. The sanction is largely symbolic and of little practical consequence, but it is another blow to the authority's reputation.

As part of the same vote, the parliament also sanctioned the European Medicines Agency and the European Environment Agency over similar issues. Jacqueline McGlade, the director of the European Environment Agency, based in Copenhagen, has been chastised for concurrently serving on the board of the Earthwatch Institute, an international environmental research and advocacy non-profit body, which has received funding from the agency. And in March, the European Medicines Agency in London was forced to place restrictions on former executive director Thomas Lönngren's employment for the next two years, following questions about his work as a pharmaceutical industry consultant.

Geslain-Lanéelle says she hopes that parliament will lift its sanction when it votes on the matter again in the autumn, after it has received a report from the European Court of Auditors on the handling of potential conflicts of interest at EFSA.

But Bánáti warns that the authority's conflict-of-interest rules risk becoming too restrictive. "It is important to understand that scientists who work for EFSA do so in an unpaid capacity, offering their expertise as a public service," she says. "EFSA should respect the free choices of all the scientists of which it has need to do its valuable work, to manage their own careers and make their own choices as they see fit." ■ SEE EDITORIAL P.279

HIGHER EDUCATION

Go West, young Russian

President Putin to back scheme for students to study abroad.

BY QUIRIN SCHIERMEIER

In an effort to grow its scientific workforce and to stimulate international research collaborations, the Russian government is set to pay for thousands of Russian students to attend top universities around the world. But to benefit from the generous scholarships, the students must agree to apply their new-found skills back home — assuming that jobs will be waiting for them when they return.

Vladimir Putin, who took office as president last week following a controversial election, is expected to officially approve the five-billion-rouble (US$165-million) Global Education programme by the end of this month. His pre-election promises included a pledge to substantially increase government funding of science and education (see *Nature* **483,** 253–254; 2012).

The programme will be run by the Strategic Initiatives Agency, a government-funded bureau set up last year with a view to promote social and economic innovation in Russia. The first call for applications should be launched next month, says Dmitry Peskov, who is head of the agency's division for young professionals and oversees the programme. "We have the means to very generously support up to 2,000 talented Russian students per year," he says.

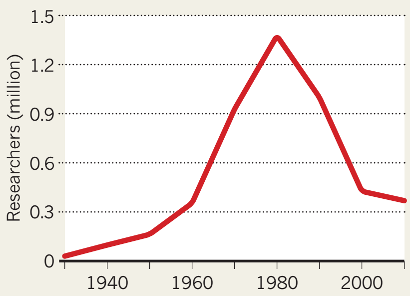
The scheme will initially operate for three years, but may be extended following a performance review planned for 2015. Students in all fields of science, technology, medicine, social science and business will be eligible for the grant — as long as they attend one of the top 300 universities in the Times Higher Education World University Rankings, says Peskov.

Students will be asked to sign a contract with the agency, in which they agree to return to Russia and secure professional work there for at least three years after graduation. If they sign up, the agency will cover their travel, tuition fees and living expenses. But they will also be obliged to pay back the full stipend if they choose not to return.

➲ NATURE.COM
For more on Russian science, see:
go.nature.com/bcgxcs

AFTER THE FALL

Russia's ministry of education and science estimates that the country now has only one-quarter of the number of researchers who were working during the prime of the Soviet Union.



"The good thing is that the initiative is in the hands of students, who will be selected — or not — on the basis of merit by foreign schools," says Konstantin Severinov, a molecular biologist at Rutgers University in Piscataway, New Jersey, who runs research groups at the Russian Academy of Sciences' institutes for molecular genetics and gene biology in Moscow. "That way, Russian university administrators cannot exert too much control" over which students receive the awards, he says.

Although he welcomes the scheme, Severinov warns that it is far from certain that there will be adequate career opportunities for the returnees. Russian science continues to struggle to regain the strength of its Soviet glory days (see 'After the fall'), and domestic high-technology industries are still in their infancy. In the short term, the lack of jobs may force these students to seek work abroad, says Severinov, contributing to the brain drain that the programme is meant to reduce.

However, similar schemes have proven effective in other countries. China — now a scientific powerhouse — has benefited considerably from government-sponsored overseas training of hundreds of thousands of students since the 1970s. The students' international experience also helps to bolster international research partnerships once they return home. A smaller programme, running since 1994, has helped to rejuvenate science in Kazakhstan. And Brazil, where scientists and engineers are in high demand, last year announced plans to send 75,000 students abroad by the end of 2014 (see *Nature* go.nature.com/x4vaoy; 2011).

Working with foreign supervisors can help to open up valuable research opportunities, says chemist Xinjiao Wang of the Ruhr University Bochum, Germany. Last year, the China Scholarship Council, based in Beijing, which funds international study, gave Wang an 'outstanding student' award worth $5,000 for her graduate research on nickel compounds at the University of Erlangen-Nürnberg, Germany. "In Germany, I've really learned how to create new ideas in science," she says. ■

CORRECTION

In the Editorial 'Price of freedom' (*Nature* **485,** 148; 2012), we stated that 'plenty of European scientists will be lost'. 'European scientists' should have been 'Europan science', as we meant to refer to science on the Jovian moon Europa.

Members of the European Parliament have also expressed their displeasure, last week voting by a narrow majority to defer approval of EFSA's 2010 budget report. The sanction is largely symbolic and of little practical consequence, but it is another blow to the authority's reputation.

As part of the same vote, the parliament also sanctioned the European Medicines Agency and the European Environment Agency over similar issues. Jacqueline McGlade, the director of the European Environment Agency, based in Copenhagen, has been chastised for concurrently serving on the board of the Earthwatch Institute, an international environmental research and advocacy non-profit body, which has received funding from the agency. And in March, the European Medicines Agency in London was forced to place restrictions on former executive director Thomas Lönngren's employment for the next two years, following questions about his work as a pharmaceutical industry consultant.

Geslain-Lanéelle says she hopes that parliament will lift its sanction when it votes on the matter again in the autumn, after it has received a report from the European Court of Auditors on the handling of potential conflicts of interest at EFSA.

But Bánáti warns that the authority's conflict-of-interest rules risk becoming too restrictive. "It is important to understand that scientists who work for EFSA do so in an unpaid capacity, offering their expertise as a public service," she says. "EFSA should respect the free choices of all the scientists of which it has need to do its valuable work, to manage their own careers and make their own choices as they see fit." ■ SEE EDITORIAL P.279

HIGHER EDUCATION

Go West, young Russian

President Putin to back scheme for students to study abroad.

BY QUIRIN SCHIERMEIER

In an effort to grow its scientific workforce and to stimulate international research collaborations, the Russian government is set to pay for thousands of Russian students to attend top universities around the world. But to benefit from the generous scholarships, the students must agree to apply their new-found skills back home — assuming that jobs will be waiting for them when they return.

Vladimir Putin, who took office as president last week following a controversial election, is expected to officially approve the five-billion-rouble (US$165-million) Global Education programme by the end of this month. His pre-election promises included a pledge to substantially increase government funding of science and education (see *Nature* **483,** 253–254; 2012).

The programme will be run by the Strategic Initiatives Agency, a government-funded bureau set up last year with a view to promote social and economic innovation in Russia. The first call for applications should be launched next month, says Dmitry Peskov, who is head of the agency's division for young professionals and oversees the programme. "We have the means to very generously support up to 2,000 talented Russian students per year," he says.

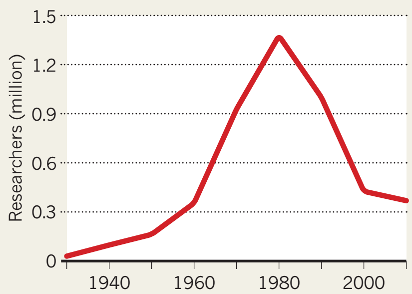
The scheme will initially operate for three years, but may be extended following a performance review planned for 2015. Students in all fields of science, technology, medicine, social science and business will be eligible for the grant — as long as they attend one of the top 300 universities in the Times Higher Education World University Rankings, says Peskov.

➲ NATURE.COM
For more on Russian science, see:
go.nature.com/bcgxcs

Students will be asked to sign a contract with the agency, in which they agree to return to Russia and secure professional work there for at least three years after graduation. If they sign up, the agency will cover their travel, tuition fees and living expenses. But they will also be obliged to pay back the full stipend if they choose not to return.

AFTER THE FALL

Russia's ministry of education and science estimates that the country now has only one-quarter of the number of researchers who were working during the prime of the Soviet Union.



"The good thing is that the initiative is in the hands of students, who will be selected — or not — on the basis of merit by foreign schools," says Konstantin Severinov, a molecular biologist at Rutgers University in Piscataway, New Jersey, who runs research groups at the Russian Academy of Sciences' institutes for molecular genetics and gene biology in Moscow. "That way, Russian university administrators cannot exert too much control" over which students receive the awards, he says.

Although he welcomes the scheme, Severinov warns that it is far from certain that there will be adequate career opportunities for the returnees. Russian science continues to struggle to regain the strength of its Soviet glory days (see 'After the fall'), and domestic high-technology industries are still in their infancy. In the short term, the lack of jobs may force these students to seek work abroad, says Severinov, contributing to the brain drain that the programme is meant to reduce.

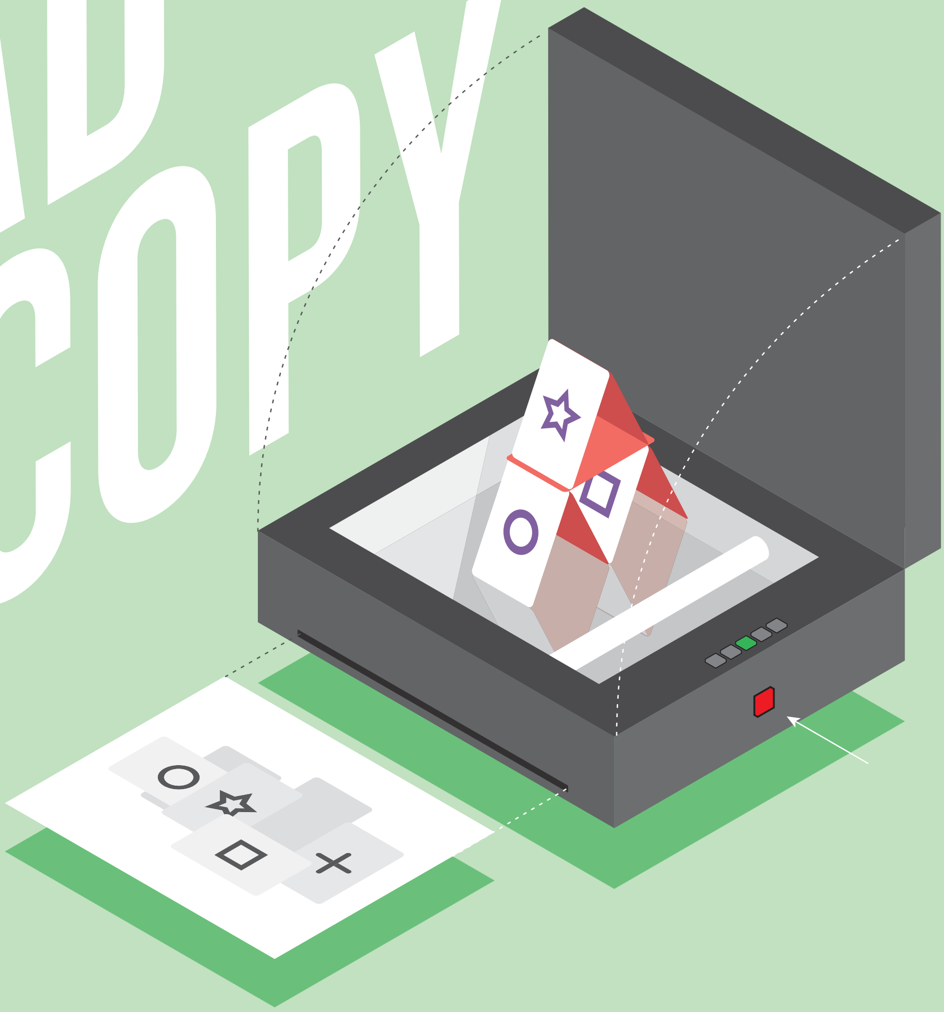
However, similar schemes have proven effective in other countries. China — now a scientific powerhouse — has benefited considerably from government-sponsored overseas training of hundreds of thousands of students since the 1970s. The students' international experience also helps to bolster international research partnerships once they return home. A smaller programme, running since 1994, has helped to rejuvenate science in Kazakhstan. And Brazil, where scientists and engineers are in high demand, last year announced plans to send 75,000 students abroad by the end of 2014 (see *Nature* go.nature.com/x4vaoy; 2011).

Working with foreign supervisors can help to open up valuable research opportunities, says chemist Xinjiao Wang of the Ruhr University Bochum, Germany. Last year, the China Scholarship Council, based in Beijing, which funds international study, gave Wang an 'outstanding student' award worth $5,000 for her graduate research on nickel compounds at the University of Erlangen-Nürnberg, Germany. "In Germany, I've really learned how to create new ideas in science," she says. ■

CORRECTION

In the Editorial 'Price of freedom' (*Nature* **485,** 148; 2012), we stated that 'plenty of European scientists will be lost'. 'European scientists' should have been 'Europan science', as we meant to refer to science on the Jovian moon Europa.

BAD COPY



JASIEK KRZYSZTOFIAK

IN THE WAKE OF HIGH-PROFILE CONTROVERSIES, PSYCHOLOGISTS ARE FACING UP TO PROBLEMS WITH REPLICATION.

BY ED YONG

For many psychologists, the clearest sign that their field was in trouble came, ironically, from a study about premonition. Daryl Bem, a social psychologist at Cornell University in Ithaca, New York, showed student volunteers 48 words and then abruptly asked them to write down as many as they could remember. Next came a practice session: students were given a random subset of the test words and were asked to type them out. Bem found that some students were more likely to remember words in the test if they had later practised them. Effect preceded cause.

Bem published his findings in the *Journal of Personality and Social Psychology* (*JPSP*) along with eight other experiments[1] providing evidence for what he refers to as "psi", or psychic effects. There is, needless to say, no shortage of scientists sceptical about his claims. Three research teams independently tried to replicate the effect Bem had reported and, when they could not, they faced serious obstacles to publishing their results. The episode served as a wake-up call. "The realization that some proportion of the findings in the literature simply might not replicate was brought home by the fact that there are more and more of these counterintuitive findings in the literature," says Eric-Jan Wagenmakers, a mathematical psychologist from the University of Amsterdam.

Positive results in psychology can behave like rumours: easy to release but hard to dispel. They dominate most journals, which strive to present new, exciting research. Meanwhile, attempts to replicate those studies, especially when the findings are negative, go unpublished, languishing in personal file drawers or circulating in conversations around the water cooler. "There are some

experiments that everyone knows don't replicate, but this knowledge doesn't get into the literature," says Wagenmakers. The publication barrier can be chilling, he adds. "I've seen students spending their entire PhD period trying to replicate a phenomenon, failing, and quitting academia because they had nothing to show for their time."

These problems occur throughout the sciences, but psychology has a number of deeply entrenched cultural norms that exacerbate them. It has become common practice, for example, to tweak experimental designs in ways that practically guarantee positive results. And once positive results are published, few researchers replicate the experiment exactly, instead carrying out 'conceptual replications' that test similar hypotheses using different methods. This practice, say critics, builds a house of cards on potentially shaky foundations.

These problems have been brought into sharp focus by some high-profile fraud cases, which many believe were able to flourish undetected because of the challenges of replication. Now psychologists are trying to fix their field. Initiatives are afoot to assess the scale of the problem and to give replication attempts a chance to be aired. "In the past six months, there are many more people talking and caring about this," says Joseph Simmons, an experimental psychologist at the University of Pennsylvania in Philadelphia. "I'm hoping it's reaching a tipping point."

PERVASIVE BIAS

Psychology is not alone in facing these problems. In a now-famous paper[2], John Ioannidis, an epidemiologist currently at Stanford School of Medicine in California argued that "most published research findings are false", according to statistical logic. In a survey of 4,600 studies from across the sciences, Daniele Fanelli, a social scientist at the University of Edinburgh, UK, found that the proportion of positive results rose by more than 22% between 1990 and 2007 (ref. 3). Psychology and psychiatry, according to other work by Fanelli[4], are the worst offenders: they are five times more likely to report a positive result than are the space sciences, which are at the other end of the spectrum (see 'Accentuate the positive'). The situation is not improving. In 1959, statistician Theodore Sterling found that 97% of the studies in four major psychology journals had reported statistically significant positive results[5]. When he repeated the analysis in 1995, nothing had changed[6].

One reason for the excess in positive results for psychology is an emphasis on "slightly freak-show-ish" results, says Chris Chambers, an experimental psychologist at Cardiff University, UK. "High-impact journals often regard psychology as a sort of parlour-trick area," he says. Results need to be exciting, eye-catching, even implausible. Simmons says that the blame lies partly in the review process. "When we review papers, we're often making authors prove that their findings are novel or interesting," he says. "We're not often making them prove that their findings are true."

Simmons should know. He recently published a tongue-in-cheek paper in *Psychological Science* 'showing' that listening to the song *When I'm Sixty-four* by the Beatles can actually reduce a listener's age by 1.5 years[7]. Simmons designed the experiments to show how "unacceptably easy" it can be to find statistically significant results to support a hypothesis. Many psychologists make on-the-fly decisions about key aspects of their studies, including how many volunteers to recruit, which variables to measure and how to analyse the results. These choices could be innocently made, but they give researchers the freedom to torture experiments and data until they produce positive results.

In a survey of more than 2,000 psychologists, Leslie John, a consumer psychologist from Harvard Business School in Boston, Massachusetts, showed that more than 50% had waited to decide whether to collect more data until they had checked the significance of their results, thereby allowing them to hold out until positive results materialize. More than 40% had selectively reported studies that "worked"[8]. On average, most respondents felt that these practices were defensible. "Many people continue to use these approaches because that is how they were taught," says Brent Roberts, a psychologist at the University of Illinois at Urbana–Champaign.

All this puts the burden of proof on those who try to replicate studies — but they face a tough slog. Consider the aftermath of Bem's notorious paper. When the three groups who failed to reproduce the word-recall results combined and submitted their results for publication, the *JPSP*, *Science* and *Psychological Science* all said that they do not publish straight replications. The *British Journal of Psychology* sent the paper out for peer review, but rejected it. Bem was one of the peer reviewers on the paper. The beleaguered paper eventually found a home at *PLoS ONE*[9], a journal that publishes all "technically sound" papers, regardless of novelty.

"I've done everything possible to encourage replications," says Bem, who stands by his results, and has put details of all his methods and tests online. But he adds that one replication paper is uninformative on its own. "It's premature," he says. "It can take years to figure out what can make a replication fail or succeed. You need a meta-analysis of many experiments."

> "TO SHOW THAT 'A' IS TRUE, YOU DON'T DO 'B'. YOU DO 'A' AGAIN."

Stéphane Doyen, a cognitive psychologist at the Free University of Brussels, encountered similar issues when he and his colleagues failed to replicate a classic experiment by John Bargh from Yale University in New Haven, Connecticut, showing that people walk more slowly if they have been unconsciously primed with age-related words[10]. After several rejections, Doyen's paper was also eventually published in *PLoS ONE*[11], and drew an irate blog post from Bargh. Bargh described Doyen's team as "inexpert researchers" and later took issue with the writer of this story for a blog post about the exchange. Bargh says that he responded so strongly partly because he saw growing scepticism of the idea that unconscious thought processes are important, and felt that damage was being done to the field.

Of course, one negative replication does not invalidate the original result. There are many mundane reasons why such attempts might not succeed. If the original effect is small, negative results will arise through chance alone. The volunteers in a replication attempt might differ from those in the original. And one team might simply lack the skill to reproduce another's experiments.

"The conduct of subtle experiments has much in common with the direction of a theatre performance," says Daniel Kahneman, a Nobel-prizewinning psychologist at Princeton University in New Jersey. Trivial details such as the day of the week or the colour of a room could affect the results, and these subtleties never make it into methods sections. Bargh argues, for example, that Doyen's team exposed its volunteers to too many age-related words, which could have drawn their attention to the experiment's hidden purpose. In priming studies, "you must tweak the situation just so, to make the manipulation strong enough to work, but not salient enough to attract even a little attention", says Kahneman. "Bargh has a knack that not all of us have." Kahneman says that he attributes a special 'knack' only to those who have found an effect that has been reproduced in hundreds of experiments. Bargh says of his priming experiments that he "never wanted there to be some secret knowledge about how to make these effects happen. We've always tried to give that knowledge away but maybe we should specify more details about how to do these things".

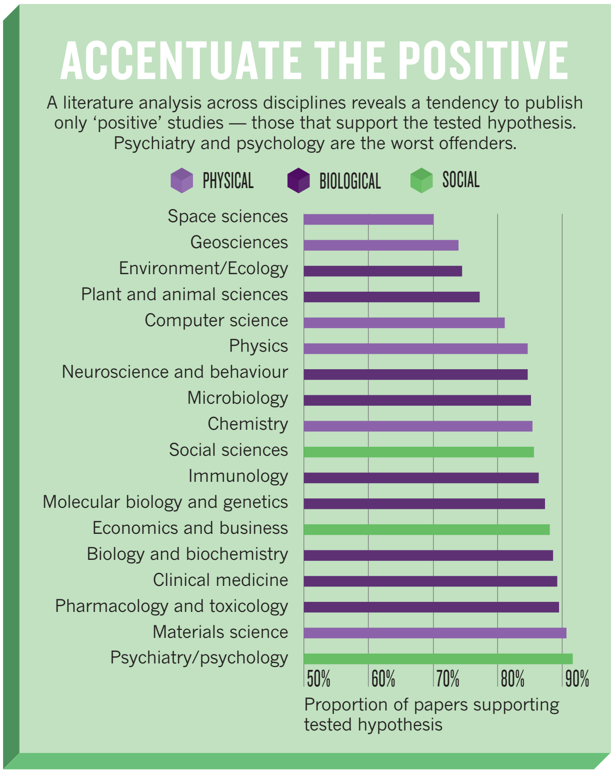
After Bargh's 1996 paper on unconscious priming, dozens of other labs followed suit with their own versions of priming experiments. Volunteers who were primed

➲ NATURE.COM
To read a World View on lessons from fraud in psychology:
go.nature.com/bkndew

by holding a heavy clipboard, for example, took interview candidates more seriously and deemed social problems to be more pressing than did those who held light boards[12]. And people primed with words relating to cleanliness judged dirty deeds more leniently[13].

Such conceptual replications are useful for psychology, which often deals with abstract concepts. "The usual way of thinking would be that [a conceptual replication] is even stronger than an exact replication. It gives better evidence for the generalizability of the effect," says Eliot Smith, a psychologist at Indiana University in Bloomington and an editor of *JPSP*.

But to other psychologists, reliance on conceptual replication is problematic. "You can't replicate a concept," says Chambers. "It's so subjective. It's anybody's guess as to how similar something needs to be to count as a conceptual replication." The practice also produces a "logical double-standard", he says. For example, if a heavy clipboard unconsciously influences people's judgements, that could be taken to conceptually replicate the slow-walking effect. But if the weight of the clipboard had no influence, no one would argue that priming had been conceptually falsified. With its ability to verify but not falsify, conceptual replication allows weak results to support one another. "It is the scientific embodiment of confirmation bias," says Brian Nosek, a social psychologist from the University of Virginia in Charlottesville. "Psychology would suffer if it wasn't practised but it doesn't replace direct replication. To show that 'A' is true, you don't do 'B'. You do 'A' again."



ACCENTUATE THE POSITIVE

A literature analysis across disciplines reveals a tendency to publish only 'positive' studies — those that support the tested hypothesis. Psychiatry and psychology are the worst offenders.

SOURCE: REF. 4

MISSED MISCONDUCT

These practices can create an environment in which misconduct goes undetected. In November 2011, Diederik Stapel, a social psychologist from Tilburg University in the Netherlands and a rising star in the field, was investigated for, and eventually confessed to, scientific fraud on a massive scale. Stapel had published a stream of sexy, attention-grabbing studies, showing for example that disordered environments, such as a messy train station, promote discrimination[14]. But all the factors making replication difficult helped him to cover his tracks. The scientific committee that investigated his case wrote, "Whereas all these excessively neat findings should have provoked thought, they were embraced … People accepted, if they even attempted to replicate the results for themselves, that they had failed because they lacked Mr Stapel's skill." It is now clear that Stapel manipulated and fabricated data in at least 30 publications.

Stapel's story mirrors those of psychologists Karen Ruggiero and Marc Hauser from Harvard University in Cambridge, Massachusetts, who published high-profile results on discrimination and morality, respectively. Ruggiero was found guilty of research fraud in 2001 and Hauser was found guilty of misconduct in 2010. Like Stapel, they were exposed by internal whistle-blowers. "If the field was truly self-correcting, why didn't we correct any single one of them?" asks Nosek.

Driven by these controversies, many psychologists are now searching for ways to encourage replications. "I think psychology has taken the lead in addressing this challenge," says Jonathan Schooler, a cognitive psychologist at the University of California, Santa Barbara. In January, Hal Pashler, a psychologist from the University of California, San Diego, in La Jolla and his colleagues created a website called PsychFileDrawer where psychologists can submit unpublished replication attempts, whether successful or not. The site has been warmly received but has only nine entries so far. There are few incentives to submit: any submission opens up scientists to criticisms from colleagues and does little to help their publication record.

Matthew Lieberman, a social psychologist from University of California, Los Angeles, suggests a different approach. "The top psychology programmes in the United States could require graduate students to replicate one of several nominated studies within their own field," he says. The students would build their skills and get valuable early publications, he says, and the field would learn whether surprising effects hold up.

Wagenmakers argues that replication attempts should also be published under different rules. Like clinical trials in medicine, he says, they should be pre-registered to avoid the post-hoc data-torturing practices that Simmons describes, and published irrespective of outcome. Engaging or even collaborating with the original authors early on could pre-empt any later quibbling over methods.

These changes may be a far-off hope. Some scientists still question whether there is a problem, and even Nosek points out that there are no solid estimates of the prevalence of false positives. To remedy that, late last year, he brought together a group of psychologists to try to reproduce every study published in three major psychological journals in 2008. The teams will adhere to the original experiments as closely as possible and try to work with the original authors. The goal is not to single out individual work, but to "get some initial evidence about the odds of replication" across the field, Nosek says.

Some researchers are agnostic about the outcome, but Pashler expects to see confirmation of his fears: that the corridor gossip about irreproducible studies and the file drawers stuffed with failed attempts at replication will turn out to be real. "Then, people won't be able to dodge it," he says. ■

Ed Yong *is a freelance writer based in London and author of the blog 'Not Exactly Rocket Science'.*

1. Bem, D. J. *J. Pers. Soc. Psych.* **100,** 407–425 (2011).
2. Ioannidis, J. P. A. *PLoS Med* **2,** e124 (2005).
3. Fanelli, D. *Scientometrics* **90,** 891–904 (2011).
4. Fanelli, D. *PLoS ONE* **5,** e10068 (2010).
5. Sterling, T. D. *J. Am. Stat. Assoc.* **54,** 30–34 (1959).
6. Sterling, T. D., Rosenbaum, W. L. & Weinkam, J. J. *Am. Stat.* **49,** 108–112 (1995).
7. Simmons, J. P., Nelson, L. D. & Simonsohn, U. *Psychol. Sci.* **22,** 1359–1366 (2011).
8. John, L. K., Loewenstein, G. & Prelec, D. *Psychol. Sci.* http://dx.doi.org/10.1177/0956797611430953 (2012).
9. Ritchie, S. J., Wiseman, R. & French, C. C. *PLoS ONE* **7,** e33423 (2012).
10. Bargh, J. A., Chen, M., Burrows, L. *J. Pers. Soc. Psych.* **71,** 230–244 (1996).
11. Doyen, S., Klein, O., Pichon, C.-L. & Cleeremans, A. *PLoS ONE* **7,** e29081 (2012).
12. Jostmann, N. B, Lakens, D. & Schubert, T. W. *Psychol. Sci.* **20,** 1169–1174 (2009).
13. Schnall, K, Benton, J. & Harvey, S. *Psychol. Sci.* **19,** 1219–1222 (2008).
14. Stapel, D. A. & Lindenberg, S. *Science* **332,** 251–253 (2011).

THE NEW TOP DOG

Shape-shifting coyotes have evolved to take advantage of a landscape transformed by people. Scientists are now discovering just how wily the creatures are.

BY SHARON LEVY

Near the dawn of time, the story goes, Coyote saved the creatures of Earth. According to the mythology of Idaho's Nez Perce people, the monster Kamiah had stalked into the region and was gobbling up the animals one by one. The crafty Coyote evaded Kamiah but didn't want to lose his friends, so he let himself be swallowed. From inside the beast, Coyote severed Kamiah's heart and freed his fellow animals. Then he chopped up Kamiah and threw the pieces to the winds, where they gave birth to the peoples of the planet.

European colonists took a very different view of the coyote (*Canis latrans*) and other predators native to North America. The settlers hunted wolves to extinction across most of the southerly 48 states. They devastated cougar and bobcat populations and attacked coyotes. But unlike the other predators, coyotes have thrived in the past 150 years. Once restricted to the western plains, they now occupy most of the continent and have invaded farms and cities, where they have expanded their diet to include squirrels, household pets and discarded fast food.

Researchers have long known the coyote as a master of adaptation, but studies over the past few years are now revealing how these unimposing relatives of wolves and dogs have managed to succeed where many other creatures have suffered. Coyotes have flourished in part by exploiting the changes that people have made to the environment, and their opportunism goes back thousands of years. In the past two centuries, coyotes have taken over part of the wolf's former ecological niche by preying on deer and even on an endangered group of caribou. Genetic studies reveal that the coyotes of northeastern America — which are bigger than their cousins elsewhere — carry wolf genes that their ancestors picked up through interbreeding. This lupine inheritance has given northeastern coyotes the ability to bring down adult deer — a feat seldom attempted by the smaller coyotes of the west.

Wolf genes make the coyotes of northeastern North America bigger and stronger than those elsewhere.

The lessons learned from coyotes can help researchers to understand how other mid-sized predators respond when larger carnivores are wiped out. In sub-Saharan Africa, for example, intense hunting of lions and leopards has led to a population explosion of olive baboons, which are now preying on smaller primates and antelope, causing a steep decline in their numbers.

Yet even among such opportunists, coyotes stand out as the champions of change. "We need to stop looking at these animals as static entities," says mammalogist Roland Kays of the North Carolina Museum of Natural Sciences in Raleigh. "They're evolving."

At a fast rate, too. Two centuries ago, coyotes led a very different life, hunting rabbits, mice and insects in the grasslands of the Great Plains. Weighing only 10 to 12 kilograms on average, they could not compete in the forests with the much larger grey wolves (*Canis lupus*), which are quick to dispatch coyotes that try to scavenge their kills.

The big break for coyotes came when settlers pushed west, wiping out the resident wolves. Coyotes could thrive because they breed more quickly than wolves and have a more varied diet. Since then, their menu has grown and so has their range; they have invaded all the mainland United States (with the exception of northern Alaska) and Mexico, as

KITCHIN AND HURST/ALL CANADA PHOTOS/CORBIS

well as large parts of southern Canada (see 'On the move').

The animals that arrived in the northeastern United States and Canada in the 1940s and 50s were significantly larger on average than those on the Great Plains, sometimes topping 16 kilograms. Kays and his colleagues studied the rapid changes in coyote physique by analysing mitochondrial DNA and skull measurements of more than 100 individuals collected in New York state and throughout New England. They found[1] that these northeastern coyotes carried genes from Great Lakes wolves, showing that the two species had interbred as the coyotes passed through that region. "Coyotes mated with wolves in the 1800s, when wolf populations were at low density because of human persecution," says Kays. In those circumstances, wolves had a hard time finding wolf mates, so they settled for coyotes.

Compared with the ancestral coyotes from the plains, the northeastern coyote–wolf hybrids have larger skulls, with more substantial anchoring points for their jaw muscles. Thanks in part to those changes, these beefy coyotes can take down larger prey; they even killed a 19-year-old female hiker in Nova Scotia in 2009. The northeastern coyotes have expanded their range five times faster than coyote populations in the southeastern United States, the members of which encountered no wolves as they journeyed east.

NEW TO THE CITY

Coyotes have even moved into Washington DC, appearing in Rock Creek Park in 2004, just a few miles from the White House. Christine Bozarth, a conservation geneticist at the Smithsonian Institution in Washington, has tracked their arrival and has shown that some of them are descended from the larger northeastern strain and carry wolf DNA[2]. Bozarth says the coyotes are there to stay. "They can adapt to any urban landscape; they'll raise their pups in drainage ditches and old pipes," she says. She hopes that the coyotes will help to control the deer, whose numbers are booming. But Kays says that coyotes have not made a significant dent in the northeast's deer population. "Coyotes fill part of the empty niche, but they don't completely replace wolves," he says.

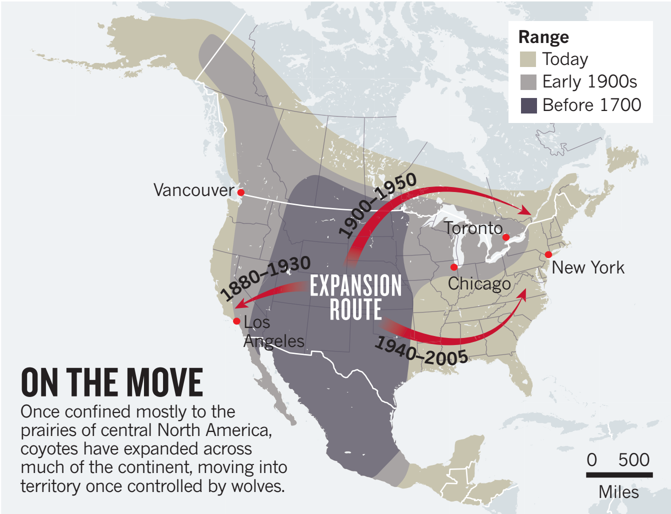
Oddly enough, it is the smaller coyotes in the southeastern United States that seem to be having a real impact on deer. About the same size as western coyotes, the southeastern ones have begun to exploit a niche left empty by the red wolves (*Canis lupus rufus*) that once roamed the southeast and specialized in hunting the region's deer, which are smaller than those in the northeast.

John Kilgo, a wildlife biologist with the US Forest Service in New Ellenton, South Carolina, and his colleagues found in a 2010 study[3] that South Carolina's deer population started to decline when coyotes arrived in the late 1980s. More recently, he and his colleagues have studied deaths among fawns, using forensic techniques right out of a murder investigation[4]. They analysed bite wounds on the carcasses and sequenced DNA in saliva left on the wounds. They also searched for scat and tracks left by the killers and noted how they had stashed uneaten remains. More than one-third of the fawn deaths were clearly caused by coyotes, and circumstantial evidence suggests that the true number might be closer to 80%. "Coyotes are acting as top predators on deer, and controlling their numbers," says Kilgo.

At first, many researchers had a hard time accepting that conclusion because they thought that coyotes were too small to affect deer populations, Kilgo says. He hopes to study how the newly arrived coyotes will affect other members of the southeastern ecosystem, including wild turkeys and predators such as raccoons, foxes and opossums.

There is no danger that the southeastern coyotes will drive the abundant deer in that region to extinction. But at the northern extreme of their range, coyotes are threatening a highly endangered band of woodland caribou (*Rangifer tarandus caribou*) in the mature forests of Quebec's Gaspésie National Park. Logging and other changes there had taken a toll on the caribou even before coyotes arrived in the region in 1973 and settled into newly cleared parts of the forest. But then coyotes started hunting caribou calves and the population dropped even further.

A 2010 study[5] found that coyotes accounted for 60% of the predation on these caribou, which now number only 140. Dominic Boisjoly, a wildlife biologist with Quebec's Ministry of Sustainable Development, Environment and Parks in Quebec City, says that the best way to protect the caribou would be to cease clear-cutting of the forest, thereby denying the predators a home.



ON THE MOVE
Once confined mostly to the prairies of central North America, coyotes have expanded across much of the continent, moving into territory once controlled by wolves.

Coyotes have been taking advantage of the changes wrought by humans for many thousands of years, according to a study of coyote fossils published this year[6]. Evolutionary biologist Julie Meachen at the National Evolutionary Synthesis Center in Durham, North Carolina, and Joshua Samuels at the John Day Fossil Beds National Monument in Kimberly, Oregon, made that discovery by measuring the size of coyote fossils dating back over the past 25,000 years. During the last ice age, coyotes were significantly larger than most of their modern counterparts and resembled the biggest of the present-day coyote–wolf hybrids in the northeast. They probably scavenged meat from kills made by dire wolves and sabre-toothed cats, and preyed on the young of the large herbivores, such as giant ground sloths, wild camels and horses, that thronged North America at that time.

But at the close of the ice age, about 13,000 years ago, most of the megafauna vanished — an extinction attributed to both climate change and the appearance of efficient Stone Age hunters. With them went the largest predators, allowing the smaller grey wolves to fill the vacant niche, which put them in competition with the largest coyotes. That conflict, as well as the loss of large herbivores, caused coyotes to shrink in stature. Within 1,000 years of the Pleistocene extinctions, coyotes had reached the same size as in most present-day populations.

Now, they're going through a whole new set of changes as they adapt to the modern landscape of North America. Genetic studies[7] show that some coyotes are even interbreeding with dogs, which could lead to a different sort of hybrid animal. Researchers are struggling to keep up with the animals and their impacts as they lope into more new regions.

"Invading a landscape emptied of wolves may trigger a whole new pathway in terms of the coyote's evolution," says Bill Ripple, an ecologist at Oregon State University in Corvallis. "And the coyote's arrival will have unpredictable effects on other species in the ecosystem." ■

Sharon Levy *is a writer based in Arcata, California, and the author of* Once and Future Giants.

1. Kays, R., Curtis, A. & Kirchman, J. J. *Biol. Lett.* **6,** 89–93 (2010).
2. Bozarth, C. A. et al. *J. Mammal.* **92,** 1070–1080 (2011).
3. Kilgo, J. C., Ray, H. C., Ruth, C. & Miller, K. V. *J. Wildlife Manage.* **74,** 929–933 (2010).
4. Kilgo, J. C. *et al. J. Wildlife Manage.* (in the press).
5. Boisjoly, D., Ouellet, J.-P. & Courtois, R. *J. Wildlife Manage.* **74,** 3–11 (2010).
6. Meachen, J. A. & Samuels, J. X. *Proc. Natl Acad. Sci. USA* **109,** 4191–4196 (2012).
7. vonHoldt, B. M. *et al. Genome Res.* **21,** 1294–1305 (2011).

COMMENT

ASTRONOMY How the transit of Venus could help to detect other worlds **p.303**

ANATOMY The nutrients and toxins found in human breast milk **p.306**

SOUND Bernie Krause on the acoustics of ecosystems **p.308**

SUSTAINABILITY India needs better management of electronic waste **p.309**

PA

The UK chief scientific adviser, John Beddington, has overseen the installation of science advisers in every department of the British government.

Beyond the great and good

Chief scientific advisers need better support and networks to ensure that science advice to governments is robust, say **Robert Doubleday** and **James Wilsdon**.

As the UK government's chief scientific adviser contemplates the end of his term of office this December, he could be forgiven a smile of satisfaction. During his five years in the post, John Beddington, a population biologist, has navigated the 2009 swine-flu outbreak and the volcanic ash cloud in 2010, limited the damage from controversies over climate science and drugs policy, and defined his own priorities — notably, his concept of a 'perfect storm' of insecurity over food, energy and water. He has also presided over the spread of 22 departmental chief scientific advisers (CSAs) into every corner of the British government, even Her Majesty's Treasury, where economics alone has traditionally held sway.

Beddington's legacy extends farther afield as a result of the international support he has built for the CSA concept. Equivalent posts have been created in New Zealand and at the European Commission in the past few years, and are proposed in Japan and at the United Nations. For those in the scientific community who have long advocated better structures for policy advice, this enthusiasm for CSAs is encouraging. But, as recruitment of Beddington's successor begins, and with similar job descriptions being rolled out elsewhere, it is timely to reflect on the strengths and weaknesses of this model.

There is no universal solution to science advice. The CSA concept has been shaped by a particularly British approach to expertise, which focuses on the credibility and character of the individual. The UK model is working well. But it has its tensions, and may not transplant easily.

> **NATURE.COM**
> C. P. Snow's portrait of science in politics:
> go.nature.com/gxskb5

A focus on the personal standing of the CSA, as in the United Kingdom, needs to be balanced by greater attention to the mix of skills, structures and staff required for high-quality policy advice. There needs to be a more open discussion by policy-makers of the trade-offs between independence and influence, and of the weight given to different disciplines and perspectives within the advisory system. Governments should do more to incorporate insight from the growing body of scholarship on science policy and expert advice. There is a need for international networks that enable science ▶

advisers from different countries to learn from one another.

The United Kingdom has had a cross-government chief scientific adviser since zoologist Solly Zuckerman's appointment in 1964. The adviser is a point of connection between science, politics and public policy: their role is to act as a personal adviser to the prime minister and the Cabinet, to lead the Government Office for Science, and to speak to the public and the media.

The US equivalent, the presidential scientific adviser, has evolved in parallel, and advisory systems from Australia and India to Ireland and Malaysia have blended elements of both models. Whereas the presidential scientific adviser serves the White House, the UK equivalent has a broad purview across all government departments. This distinction has sharpened as each department has appointed its own CSA, a process that began in 2002 in response to perceived failings in policy advice during the mid-1990s BSE crisis. By contrast, the reliance on hundreds of expert committees across the US political system has, in effect, restricted the remit of the presidential scientific adviser[1].

INFLUENTIAL MODEL

The UK model is increasingly seen as a template elsewhere. Medical scientist Peter Gluckman, appointed in 2009 as New Zealand's first CSA, plans to install scientific advisers in a number of the country's ministries. In December 2011, José Manuel Barroso, president of the European Commission, announced molecular biologist Anne Glover, former CSA to the Scottish government, as Europe's first CSA.

In Japan, the government has set out plans to install chief science advisers to the prime minister and other ministers, as part of an overhaul of the science advisory system proposed after last year's earthquake, tsunami and nuclear disaster. At the United Nations in New York, Secretary-General Ban Ki-moon announced in April that he will appoint a CSA as part of a package of reforms designed to strengthen the international governance of sustainable development, recommended ahead of the Rio+20 summit in June 2012 (ref. 2).

With so many nations adopting the CSA model, it is worth noting that the system is a product of one political culture. For example, a 2012 report by the UK House of Lords Science and Technology Committee highlighted "standing and authority within the scientific community" as the foremost essential characteristic of a CSA[3]. Therefore, they should be externally appointed senior figures, drawn from academia or industry, who can access a wide range of expertise by dint of their position and reputation.

These recommendations sit comfortably within Britain's political culture, which has long emphasized the authority of individual experts with a track record of public service. But there is more to effective science policy than being one of the great and good. Even though most UK CSAs would pass the Lords' 'scientific standing' criterion with flying colours, the report acknowledges the uneven performance of some. The variable budgets across different departments may be partly responsible. The Lords recommend increasing the CSA's status, autonomy and funding, but other factors deserve attention, such as how CSAs fit within wider policy-making structures.

"The theory and analysis of scientific advice needs to better inform its practice."

The report lists other qualities that CSAs should have: an ability to engage with stakeholders; to manage multidisciplinary teams; to understand the policy environment; and to be able to evaluate conflicting evidence from a range of perspectives. However, these are listed as only secondary considerations.

Linked to this is the vexed question of whose expertise counts. As its network of CSAs has grown, the United Kingdom has drawn advisers from across the natural sciences, and recently engineering, economics and social science. The Lords' report calls for a new post of chief social-science adviser to provide a better conduit for advice from that community. But creating stand-alone structures for different disciplines is a clumsy solution. The issue is how to integrate an appropriate mix of expert advice across government to address inherently interdisciplinary issues — from climate change to food security and obesity.

GLOBAL LESSONS

So what lessons can other countries draw from the UK system? First, although a focus on the credentials and character of individual CSAs is important, it needs to be balanced with the mix of skills, structures and staff that is essential for high-quality scientific advice. Science-policy expert Roger Pielke Jr has argued that it is a mistake to base advisory systems on the idea of a CSA as "a heroic individual with special influence on policy making" (see go.nature.com/broi9t). Neither should CSAs be seen as a 'one-way membrane' that allows science into politics while protecting science from political influence; governments need to create processes for integrating the two.

Second, there needs to be greater recognition of the contribution that different disciplines and perspectives make to an effective advisory system — including the social sciences, arts and humanities, but also civil society and the public. James Killian, for example, who advised US President Dwight D. Eisenhower and was widely regarded as the most successful US science adviser, was not a scientist and had a degree in management.

Advisory systems must acknowledge the importance of politics in shaping what counts as evidence and authority, as well as the plural, conditional nature of knowledge. As noted by Andy Stirling, a science-policy researcher at the University of Sussex in Brighton, UK, expert advice is often expected to provide one interpretation and to reduce uncertainties to measurable risks. However, this process of 'closing down' limits the scope of advice[4].

Third, the theory and analysis of scientific advice needs to better inform its practice. There is now a wealth of empirical research into how advisory processes operate; a recent exercise led by the Centre for Science and Policy at the University of Cambridge, UK, sought to identify the most pressing questions about the relationship between science and policy[5]. Sheila Jasanoff, professor of science and technology studies at Harvard Kennedy School of Government in Cambridge, Massachusetts, argues that "good science in public decision-making cannot be divorced from deeper reflection on the ways in which democracies should reason"[6]. We think that CSAs would benefit from processes of learning and reflection that are more systematic.

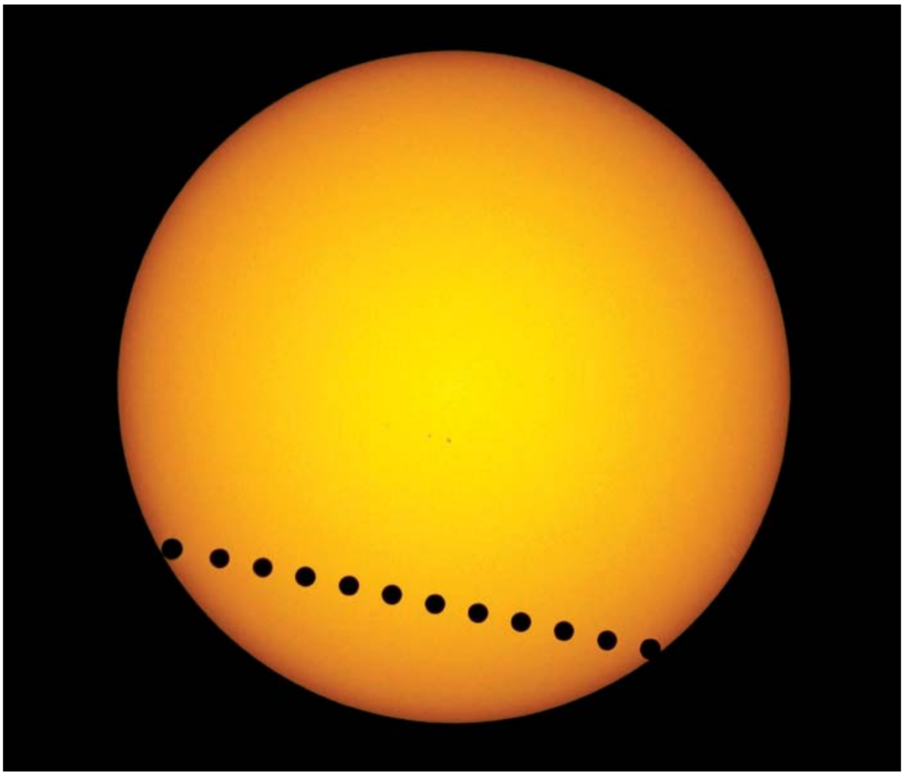
Finally, stronger international networks are required for CSAs to exchange ideas. The main forum is currently the Carnegie Group of Science Advisers, which was established in 1991 to enable CSAs and science ministers from the G8 nations to meet annually, and which has recently expanded to include Brazil, China, India, Mexico and South Africa. A more open global network is now required. In a welcome step towards this goal, the CSA of Quebec recently invited all CSAs to a meeting in Montreal in October 2012.

The learning needs to flow in all directions. As the process for appointing Beddington's successor gets under way, these lessons should be applied in London, too. ■

Robert Doubleday *is head of research at the Centre for Science and Policy, University of Cambridge, UK.* **James Wilsdon** *is professor of science and democracy at the Science Policy Research Unit, University of Sussex, Brighton, UK.*
e-mail: j.wilsdon@sussex.ac.uk

1. Pielke Jr, R. & Klein, R. A. (eds) *Presidential Science Advisors* (Springer, 2010).
2. United Nations Secretary-General's High-Level Panel on Global Sustainability *Resilient People, Resilient Planet: A Future Worth Choosing* (United Nations, 2012).
3. House of Lords Select Committee on Science and Technology *The Role And Functions Of Departmental Chief Scientific Advisers* (The Stationery Office, 2012).
4. Stirling, A. *Nature* **468,** 1029–1031 (2010).
5. Sutherland, W. J. *et al. PLoS ONE* **7,** e31824 (2012).
6. Jasanoff, S. In *The Politics of Scientific Advice* (eds Lentsch, J. & Weingart, P.) Part I, Ch. 2 (Cambridge University Press, 2011).

E. SLAWIK/SPL



A composite image of the June 2004 transit of Venus as seen from Waldenburg, Germany.

Last chance to see

The June 2012 transit of Venus across the Sun offers an opportunity to check our methods for spotting distant planets crossing far-away stars, says **Jay M. Pasachoff**.

The sight of Venus silhouetted against the Sun is exceedingly rare. Since 1631, when a transit of Venus was predicted but unobservable from much of Europe (where most early telescopes were located), such transits have been seen only six times: in 1639, 1761, 1769, 1874, 1882 and 2004. Our last chance, until 2117, to see and study such a transit from Earth will be on 5–6 June (http://transitofvenus.info).

Earlier transits were subject to intense scientific scrutiny. Hundreds of expeditions were sent out from around the world to observe the transits of 1761 and 1769, in efforts to triangulate the distance from Earth to the Sun. Today, the distance known as the astronomical unit has long been settled by more accurate means. Many people now think of transits within our Solar System only as public-outreach opportunities. Yet there is still much to be learned from modern transits, by taking advantage of new ideas, techniques and scientific capabilities.

For example, my colleagues and I used spacecraft observations of Mercury's 1999 transit to settle a long-standing mystery about the appearance of transiting planets.

Despite modern scientists' ability to send probes to Venus for close-up scrutiny, its transits as seen from Earth (and elsewhere) still provide unique information, and they give us the opportunity to calibrate or improve our methods for finding far-off planets around distant stars.

BLACK DROP

I became interested in transits a decade ago, when I learned from a historical talk that most contemporary books and articles wrongly attribute the cause of the phenomenon known as the black-drop effect[1].

In the 1700s, the method for determining the average distance to the Sun depended on the accurate measurement, to about a second, of the period from when Venus fully entered the solar disk to when it began to depart. But such measures were confounded by the appearance of a dark ligature between Venus's silhouette and the space outside the solar disk. This 'black drop' would grow for about a minute and then pop. The effect made the uncertainty in timing much worse than had been hoped for. For centuries, the black-drop effect limited the solution to what was arguably the most important problem in astronomy: the distances of planets from the Sun. Even by the 2000s, most publications still attributed the black-drop effect to the diffraction of light around Venus, the refraction of light by its atmosphere or even an optical illusion — even though the evidence showed these explanations to be false[1].

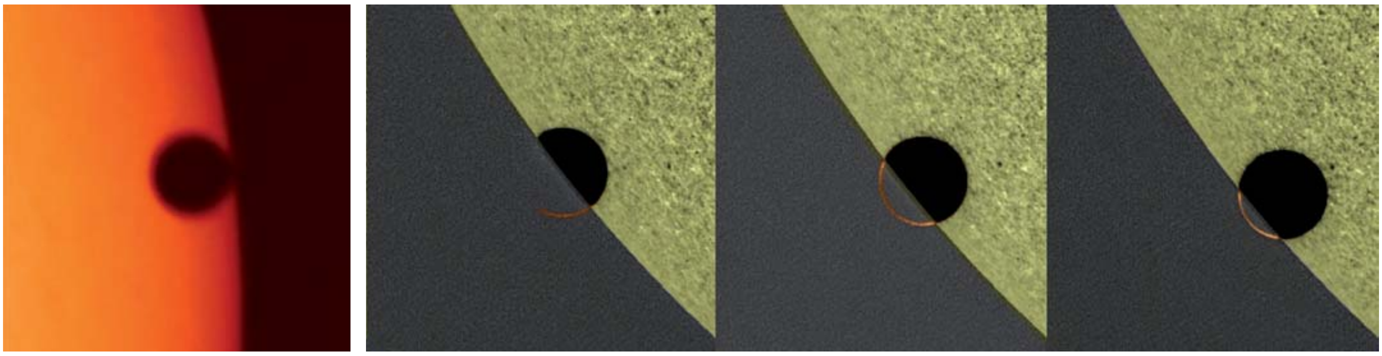
➲ NATURE.COM
For more on exoplanet astronomy, see: go.nature.com/msrrqh

Glenn Schneider, an astronomer at the University of Arizona in Tucson, and I looked for a black drop in the data taken for Mercury's 1999 transit by NASA's TRACE satellite. We found one, confirming that neither the planet's nor Earth's atmosphere was needed to produce the phenomenon (although Earth's atmosphere exaggerates the effect when viewed from Earth's surface). Our analysis showed that two effects could fully explain the black drop as seen from space: the inherent blurriness of the image caused by the finite size of the telescope, and an extreme dimming of the Sun's surface just inside its apparent outer edge[2,3]. Apart from settling a point of historic interest, understanding this effect may yet help to unravel observations of other planets' transits across far-away stars.

My colleagues, my students and I set out to test our ideas about the black-drop effect through observations of the 8 June 2004 transit of Venus. This brought an unrelated surprise. Images taken by TRACE during ingress and egress showed a rim of light that appeared over the trailing side of Venus about 20 minutes before it fully entered the solar disk[4] — the result of refraction of sunlight towards us by Venus's atmosphere. We were flabbergasted that Venus's atmosphere was so visible (although it turns out that this effect had also been noted during the 1874 transit). The arc of light was asymmetrical — brighter at some latitudes on the planet than at others.

With colleagues, we later compared our results to data from land-based telescopes[5] and from the European Space Agency's Venus Express probe, which arrived at the planet in 2006 (ref. 4). These comparisons helped to confirm, for example, observations that the planet's haze is at lower altitudes at its poles than at its equator.

The transit showed us some things that the probe could not. For many atmospheric measures, Venus Express sees only one small slice of the atmosphere at a time during sunrise or sunset: a measure is taken at one latitude one day and at another latitude the next, making it impossible to know whether any differences are due to changes in space or to changes in time. Only through



WILLIAMS COLLEGE EXPEDITION

J. M. PASACHOFF (WILLIAMS COLLEGE) AND G. SCHNEIDER (U. ARIZONA)/LMSAL AND NASA

Left: the black-drop effect connects the transiting planet to dark space. Right: Venus's atmosphere seen from space as refracted sunlight, 2004 transit.

transit observations can we see an entire arc of the planet's atmosphere at once. Further comparisons between the two will help to cross-calibrate both techniques.

From that same 2004 transit, we were able to use NASA's ACRIM III instrument on the ACRIMSAT spacecraft to detect not just the 0.1% drop in light from the Sun during the central part of Venus's passage, but also the smaller changes that resulted when Venus blocked parts of the dim outer regions of the solar disk[6]. We could even calculate the diameter of Venus (which is already known), which should help to gauge the accuracy of this technique in assessing the sizes of exoplanets.

HOME TRUTHS

The view that ACRIM III and the Total Irradiance Monitor on NASA's SORCE spacecraft will have of Venus's transit across the Sun is analogous to the view that planet-hunting satellites and ground-based telescopes have of exoplanets moving across other stars. NASA's Kepler spacecraft has found more than 2,000 exoplanet candidates so far, but it has confirmed fewer than 100. One problem in confirming sightings is knowing whether the dip in light seen from a distant star is due to a transiting planet or to some other effect, such as starspots (akin to sunspots).

The upcoming transit offers a particularly good opportunity to study this effect. The 2004 Venus transit happened at a time of minimal solar activity, when no sunspots were visible. By contrast, a 2006 transit of Mercury — whose disk in transit is just 3% of the area of Venus as seen from Earth — happened during the passage of a large sunspot across the face of the Sun. The dip in the total solar irradiance from that transit was lost amid the noise of solar activity and the limitations of the spacecraft. As expected at this phase of the sunspot cycle, 2012 is turning out to be a year of relatively high solar activity, so we expect the Venus transit to provide a situation that more closely resembles that of 'spotty' stars hosting exoplanets.

It is too soon to know exactly how the study of transits in our Solar System will help us to interpret observations of distant exoplanets, but transits are so rare that to squander these opportunities would be a crime. We owe it to future astronomers — especially those who will observe the next transit of Venus, in 2117 — to collect as much data as possible. One never knows what will prove vital to future research.

In the eight years since the last Venus transit, the equipment has improved. The TRACE satellite has been replaced by the Solar Dynamics Observatory, which has about the same resolution, but over the entire Sun, giving a full set of filtered images every 10 seconds. A US telescope on Japan's Hinode satellite has even better resolution.

All hands will be on deck to watch the coming transit. Using telescopes on Haleakala in Hawaii, my team will look, in part, at the polarization of sunlight by Venus's atmosphere, which will tell us something about particle size. With colleagues at several institutions in France, we will use a set of nine coronagraphs around the world to study the bright arc of Venus's atmosphere.

"We owe it to the astronomers who will observe the 2117 transit of Venus to collect as much data as possible."

In the continental United States, our group will observe the transit using a massive spectrograph at the Richard B. Dunn Solar Telescope of the National Solar Observatory at Sacramento Peak in New Mexico. We have purchased a new filter at the wavelength of carbon dioxide — a major constituent of Venus's atmosphere — for these observations. This will provide a unique, detailed spectrographic study of a relatively well-known atmosphere during a transit, which we can compare to studies of unknown exoplanet atmospheres.

There are some more cunning, indirect ways of watching the transit. The Hubble Space Telescope, which is too sensitive to be pointed directly at the Sun, will be used to observe sunlight reflected from the lunar surface. The low light in this observation mimics what night-time telescopes see when studying exoplanet transits.

We have also applied for time on Hubble to observe other transits, including a 20 September 2012 transit of Venus as seen from sunlight reflected off Jupiter; if granted that time, we will also apply to observe a 5 January 2014 transit of Earth as seen from Jupiter. Many people will be keen to see what a habitable, populated planet looks like in transit, and it would be a sheer delight to watch Earth pass in front of the Sun. The next such opportunity will not be until 2026.

NASA's Cassini spacecraft, now in orbit in the Saturn system, also has an unusual vantage for spying on transits — and it can directly observe the Sun. We have already arranged for it to observe a transit of Venus on 21 December 2012.

In 1874, astronomer Richard Proctor wrote, in his book *Transits of Venus*, "Let it be hoped that the success of operations conducted by the various scientific nations in 1874 and 1882 may be such that preliminary difficulties will hereafter be remembered only as obstacles successfully removed and in good time." One can only hope that future scientists will look back at the relatively feeble attempts of twenty-first-century astronomers and say that we, too, did the best we could. ■ SEE BOOKS & ARTS P.305

Jay M. Pasachoff *is Field Memorial Professor of Astronomy at Williams College, Williamstown, Massachusetts 02167, USA. He is also chair of the International Astronomical Union's Working Group on Solar Eclipses and vice-chair of the American Astronomical Society's Historical Astronomy Division.*
e-mail: eclipse@williams.edu

1. Schaefer, B. E. *Bull. Am. Astron. Soc.* **32,** 1383 (2000).
2. Pasachoff, J. M., Schneider, G. & Golub, L. in *Transits of Venus: New Views of the Solar System and Galaxy* (ed. Kurtz, D. W.) 242–253 (Cambridge Univ. Press, 2005).
3. Schneider, G., Pasachoff, J. M. & Golub, L. *Icarus* **168,** 249–256 (2004).
4. Pasachoff, J. M., Schneider, G. & Widemann, T. *Astron. J.* **141,** 112 (2011).
5. Tanga, P. *et al. Icarus* **218,** 207–219 (2012).
6. Schneider, G., Pasachoff, J. M. & Willson, R. C. *Astrophys. J.* **641,** 565–571 (2006).

Captain James Cook's first voyage to Tahiti was one of many expeditions to use the 1769 transit of Venus to measure the distance from Earth to the Sun.

CAPTAIN G. TOBIN (1768–1838)/MITCHELL LIB. STATE LIB. OF NSW/THE BRIDGEMAN ART LIB.

ASTRONOMY

On the track of the transit

Owen Gingerich enjoys two histories of the expeditions that aimed to measure the passage of Venus across the face of the Sun.

The bright planet Venus put on a dazzling evening show when it passed Jupiter in March, but it will soon drop into the sunset. The transition from evening to morning star — which occurs when the planet overtakes Earth, every 584 days on average — usually goes unnoticed, the planet running just above or below the solar disk. But this year it will transit right across the face of the Sun, visible on 5 June in the United States (including Alaska and Hawaii) and on 6 June on the other side of the International Date Line (including in Europe). This rare phenomenon will not be repeated until 2117.

A transit of Venus is visible to the naked eye, but staring at the Sun without a filter is not recommended. Such transits passed unobserved until 1639, when the tables of planetary motion became good enough for Englishman Jeremiah Horrocks to anticipate and view one. In 1716, the astronomer Edmond Halley pointed out that transits of Venus simultaneously observed from far-flung points on Earth could be used to determine the distance to the Sun, which at the time had been only roughly guessed.

Transits of Venus come in pairs, 8 years apart, at intervals of more than a century, so the earliest opportunities to test Halley's idea came in 1761 and 1769. The eighteenth century was an age of exploration, and countries vied with each other to send expeditions to remote parts of Earth to capture the essential astronomical data — and to see what else could be found. Nowadays, the distance to the Sun has been securely established by other methods (including radar), but the transits remain rare touristic occasions, and writing about their history has become a cottage industry. Two excellent accounts are among this year's yield, both concentrating on the heroic eighteenth-century expeditions.

Chasing Venus: The Race to Measure the Heavens
ANDREA WULF
Knopf/William Heinemann: 2012. 336 pp. $26.95/£18.99

The Day the World Discovered the Sun: An Extraordinary Story of Scientific Adventure and the Race to Track the Transit of Venus
MARK ANDERSON
Da Capo: 2012. 304 pp. $26, £17.99

Historian Andrea Wulf's *Chasing Venus* is beautifully paced, alternating between expeditions, with lush descriptions of the often arduous journeys involved. She describes each group's experiences in the climactic days of the transits, some meeting disappointment as clouds ruined their pursuits. Perhaps no story was more frustrating than that of French astronomer Guillaume Le Gentil. He had intended to view the 1761 transit from India, but the English had captured the port of his destination. During the transit he found himself on the high seas without the use of his pendulum clock or an established location, so his observations were useless.

Le Gentil stayed in Asia for 8 years to wait for the second transit of the pair, exploring as far as Manila until the French Academy of Sciences ordered him back to India for the 1769 conjunction. The weather was perfect until the day of the transit, and then clouds appeared. To rub salt into the wound, the sky was clear in Manila.

Journalist Mark Anderson's arresting *The Day the World Discovered the Sun* begins with the 1761 transit, but concentrates on the three most significant journeys of the 1769 event. These were Captain James Cook's voyage to Tahiti; the Hungarian Jesuit Maximilian Hell's frigid journey to Vardø, above the Arctic Circle in Norway; and French astronomer Jean-Baptiste Chappe d'Auteroche's sweaty and insect-ridden expedition to San José del Cabo in Baja California, present-day Mexico. Anderson serves up a rich broth ▸

of details — such as that British sailors did not have soap in their rations until the 1780s, or that Cook's small ship *Endeavour* had more than 90 people on board, in part because it was expected that half the crew on a round-the-world trip would die of scurvy. (In the event, Cook engaged in a medical experiment with a diet of sauerkraut for the crew, and not a single sailor was lost to the condition.)

Both Wulf and Anderson give much attention to Chappe, the only observer to time the entrance and exit of Venus on both transits. Chappe wrote vivid and extensive travel notes, which both authors use to great effect. His wide-ranging interests would have made him, thinks Anderson, the French Benjamin Franklin. Alas, in a scene drenched with pathos, Chappe died of typhus within two weeks of writing his last journal entry in Baja California.

Unfortunately, neither book explains in simple terms why the astronomers were so keen to record to the second when Venus entered and exited the solar disk. They were triangulating the distance to the Sun with long skinny triangles, the base being the separation of the stations on Earth — which is why it was crucial to know the terrestrial coordinates of the stations.

Venus provided a reference point by which apparent positions on the Sun's face could be measured from two different locations. The duration of the passage gave the length of the path across the Sun, which could then be fitted uniquely onto the observed solar disk. The angular separation of the apparent lines of transit as seen from two different stations, plus the relative distances of Venus and Earth from the Sun and the distance between the two stations, then yielded the distance to the Sun. The numbers from the three principal stations (Tahiti, Vardø and San José del Cabo) gave a mean distance within 1% of the 149,598,000 kilometres accepted today.

The eighteenth-century efforts to track the transit helped to establish the distance to the Sun, but the accuracy was far from what astronomers had hoped for. The results of the campaigns to track the next pair of transits, in 1874 and 1882, were better but still ambiguous. Yet these simultaneously competitive and cooperative efforts set the international stage for our now-accurate measure of the solar distance — the baseline from which all cosmic distances, and ultimately the age of the Universe, are reckoned. ■ SEE COMMENT P.303

Owen Gingerich *is an astronomer and historian of science at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts.*
e-mail: ginger@cfa.harvard.edu

WESTEND61/REX FEATURES

Breast milk aided the evolution of the large human brain — but it can contain toxins.

BIOLOGY

Mammary chronicles

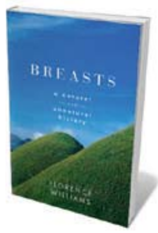
Josie Glausiusz celebrates an environmental history of the human breast.

The breast looms large in human culture and biology. The essential proteins and long-chain fatty acids in breast milk help to build babies' big brains, and the cornucopia of other components, such as virus-slaying macrophages and oligosaccharides that feed beneficial bacteria in the baby's gut, offer crucial immune protection. Unfortunately, breast milk can also contain pesticides, mercury, benzene and minuscule amounts of paint thinners, dry-cleaning fluids, rocket fuel and flame retardants.

This contaminant-crammed elixir is uniquely modern, as Florence Williams details in *Breasts*. This is no salacious tell-all, but a lively, absorbing, meticulously researched book covering all aspects of breasts, from anatomy to their role in evolution, attraction and infant bonding; changes during puberty, pregnancy and cancer; and Western society's passion for flaunting, grading and inflating them. At heart, however, the book is an environmental history of "how our breasts went from being honed by the environment to being harmed by it".

Williams, a US science journalist, uses her own body as a research tool. She sends her milk to Germany to be tested for flame retardants, delves into her family history of breast cancer and visits a suave Texas surgeon for advice on silicone breast implants. To mimic a study on early puberty, she and her seven-year-old daughter, Annabel, valiantly try to give up plastic-wrapped food as well as products containing endocrine-disrupting phthalates — including Williams's car.

Intimate explorations of breast biology have a distinguished history. In 1840, British surgeon Astley Cooper published *The Anatomy and Diseases of the Breast*, in which he observed — after injecting dyes into more than 200 disembodied breasts — that blood is transformed into milk in grape-like lobules, inside tissue cavities called alveoli. The milk then enters a network of lobes that empty into 12 or so orifices in the nipple.



Breasts: A Natural and Unnatural History
FLORENCE WILLIAMS
Norton: 2012. 352 pp. $25.95, £16.99)

Unlike any other organ, human breasts do most of their development well after birth. These plump orbs are also unique in that no other primate is so endowed: females of other species develop swellings only during lactation. Evolutionary biologists have devised elaborate stories to explain the permanent adult presence of human breasts; the most popular is that they are an adornment, like a peacock's train, for attracting the opposite sex. Williams leans more towards the ideas of anthropologist Frances Mascia-Lees, who posits that breasts' ever-present fat reserves are easily mobilized during lactation to keep pace with the baby's rapidly growing brain.

The immune support offered by human breast milk is formidable. The average new mother produces roughly 454 grams of milk from each breast every 24 hours. This elixir is not unlike cultured yoghurt, carrying 100–600 species of live bacteria, most new to science. (Mysteriously, the US National

of details — such as that British sailors did not have soap in their rations until the 1780s, or that Cook's small ship *Endeavour* had more than 90 people on board, in part because it was expected that half the crew on a round-the-world trip would die of scurvy. (In the event, Cook engaged in a medical experiment with a diet of sauerkraut for the crew, and not a single sailor was lost to the condition.)

Both Wulf and Anderson give much attention to Chappe, the only observer to time the entrance and exit of Venus on both transits. Chappe wrote vivid and extensive travel notes, which both authors use to great effect. His wide-ranging interests would have made him, thinks Anderson, the French Benjamin Franklin. Alas, in a scene drenched with pathos, Chappe died of typhus within two weeks of writing his last journal entry in Baja California.

Unfortunately, neither book explains in simple terms why the astronomers were so keen to record to the second when Venus entered and exited the solar disk. They were triangulating the distance to the Sun with long skinny triangles, the base being the separation of the stations on Earth — which is why it was crucial to know the terrestrial coordinates of the stations.

Venus provided a reference point by which apparent positions on the Sun's face could be measured from two different locations. The duration of the passage gave the length of the path across the Sun, which could then be fitted uniquely onto the observed solar disk. The angular separation of the apparent lines of transit as seen from two different stations, plus the relative distances of Venus and Earth from the Sun and the distance between the two stations, then yielded the distance to the Sun. The numbers from the three principal stations (Tahiti, Vardø and San José del Cabo) gave a mean distance within 1% of the 149,598,000 kilometres accepted today.

The eighteenth-century efforts to track the transit helped to establish the distance to the Sun, but the accuracy was far from what astronomers had hoped for. The results of the campaigns to track the next pair of transits, in 1874 and 1882, were better but still ambiguous. Yet these simultaneously competitive and cooperative efforts set the international stage for our now-accurate measure of the solar distance — the baseline from which all cosmic distances, and ultimately the age of the Universe, are reckoned. ■ SEE COMMENT P.303

Owen Gingerich *is an astronomer and historian of science at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts.*
e-mail: ginger@cfa.harvard.edu

WESTEND61/REX FEATURES

Breast milk aided the evolution of the large human brain — but it can contain toxins.

BIOLOGY

Mammary chronicles

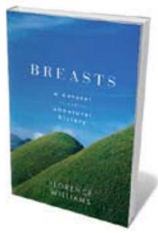
Josie Glausiusz celebrates an environmental history of the human breast.

The breast looms large in human culture and biology. The essential proteins and long-chain fatty acids in breast milk help to build babies' big brains, and the cornucopia of other components, such as virus-slaying macrophages and oligosaccharides that feed beneficial bacteria in the baby's gut, offer crucial immune protection. Unfortunately, breast milk can also contain pesticides, mercury, benzene and minuscule amounts of paint thinners, dry-cleaning fluids, rocket fuel and flame retardants.

This contaminant-crammed elixir is uniquely modern, as Florence Williams details in *Breasts*. This is no salacious tell-all, but a lively, absorbing, meticulously researched book covering all aspects of breasts, from anatomy to their role in evolution, attraction and infant bonding; changes during puberty, pregnancy and cancer; and Western society's passion for flaunting, grading and inflating them. At heart, however, the book is an environmental history of "how our breasts went from being honed by the environment to being harmed by it".

Williams, a US science journalist, uses her own body as a research tool. She sends her milk to Germany to be tested for flame retardants, delves into her family history of breast cancer and visits a suave Texas surgeon for advice on silicone breast implants. To mimic a study on early puberty, she and her seven-year-old daughter, Annabel, valiantly try to give up plastic-wrapped food as well as products containing endocrine-disrupting phthalates — including Williams's car.

Intimate explorations of breast biology have a distinguished history. In 1840, British surgeon Astley Cooper published *The Anatomy and Diseases of the Breast*, in which he observed — after injecting dyes into more than 200 disembodied breasts — that blood is transformed into milk in grape-like lobules, inside tissue cavities called alveoli. The milk then enters a network of lobes that empty into 12 or so orifices in the nipple.



Breasts: A Natural and Unnatural History
FLORENCE WILLIAMS
Norton: 2012. 352 pp. $25.95, £16.99)

Unlike any other organ, human breasts do most of their development well after birth. These plump orbs are also unique in that no other primate is so endowed: females of other species develop swellings only during lactation. Evolutionary biologists have devised elaborate stories to explain the permanent adult presence of human breasts; the most popular is that they are an adornment, like a peacock's train, for attracting the opposite sex. Williams leans more towards the ideas of anthropologist Frances Mascia-Lees, who posits that breasts' ever-present fat reserves are easily mobilized during lactation to keep pace with the baby's rapidly growing brain.

The immune support offered by human breast milk is formidable. The average new mother produces roughly 454 grams of milk from each breast every 24 hours. This elixir is not unlike cultured yoghurt, carrying 100–600 species of live bacteria, most new to science. (Mysteriously, the US National

Institutes of Health's Human Microbiome Project, which is decoding the genes of microbes from every major human surface or orifice, is not analysing breast milk.) One theory suggests that the bacteria in breast milk work as a kind of gut vaccine.

The stuff is so beneficial that companies that produce substitute breast milk are racing to replicate its ingredients, with little success so far. But Williams's investigation suggests that there may be good reasons to hope that they succeed: breast-milk toxins can include "the mercury in last week's sushi, the benzene from your gas station, … the chromium from your nearby smoke stack". Moreover, even a tiny dose of these contaminants can be harmful to babies, and such toxins have been implicated in low intelligence and cancer.

Williams reports that flame retardants, found in sofas, nursing pillows and infant car seats, can impede brain growth and affect thyroid hormones. When Williams's milk is tested for polybrominated diphenyl ethers (PBDEs), she learns that her levels are slightly above average for US women — and notes that mothers offload about 30% of their PBDE burden onto their babies if they nurse for a year. Williams breast-fed her two children for 18 months each.

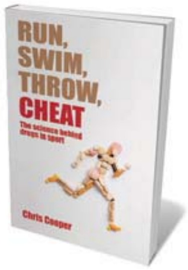
It is not just infant development that may be affected: the author describes how hormone-disrupting phthalates and bisphenol A (BPA) may be advancing puberty in girls by prematurely switching on oestrogen receptors in breast tissues. Williams looks at possible reasons behind the rise in breast cancer — globally, the leading cause of cancer-related death in women, with 1 million diagnosed each year — including better detection, hormone-replacement therapy and exposure to untested chemicals.

In one alarming account, she reports on an epidemic of male breast cancer among US marines at Camp Lejeune in North Carolina. Over three decades, starting in the 1950s, fuel tanks leaked more than 3.8 million litres of petrol into the base's groundwater. One well, which supplied drinking water to 8,000 people, contained 76 times the legal limit of benzene, a known human carcinogen.

As Williams points out, breast milk boosted brain size in our ancestors, but those brains have helped us to change the environment — which, in turn, is channelling to infant brains toxins that may impede their development. There is hope for the future, however: in 2004, the United Nations implemented the Stockholm Convention on Persistent Organic Pollutants, in which 177 countries have agreed to ban or restrict such chemicals, including some PBDEs. The United States has yet to ratify the treaty. ■

Josie Glausiusz *is a science journalist based in New York City.*
e-mail: josiegz@gmail.com

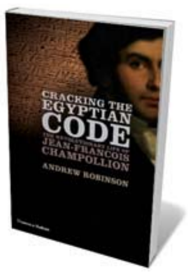
Books in brief



Run, Swim, Throw, Cheat: The Science Behind Drugs in Sport
Chris Cooper OXFORD UNIV. PRESS *288 pp. £16.99 (2012)*
Whether sprinting, swimming, lifting or leaping, elite athletes in action are phenomenal — and, as biochemist and sports scientist Chris Cooper shows in this pacy account, some are also assisted by performance-enhancing drugs. To understand a problem that is unlikely to disappear from sport completely, Cooper lays out research on the substances in question, how they work, which are illegal and the methods for detecting them. He explores a number of contexts, ranging from sexual dimorphism and the need for oxygen and key nutrients to gene doping and the science behind the tests.

Digital Vertigo: How Today's Online Social Revolution Is Dividing, Diminishing, and Disorienting Us
Andrew Keen ST MARTIN'S *256 pp. $25.99 (2012)*
In a world gripped by digital utopianism, Silicon Valley insider Andrew Keen is an uber-sceptic. Those in online communities are, says Keen, besotted with a corpse, much like James Stewart's character in Alfred Hitchcock's 1958 film *Vertigo*. Rather than offering a vast, benign e-neighbourhood, he argues, forms of social media breach privacy, encourage narcissism and promote commodification of personality. Aloneness, he says, is a necessary antidote to the hypervisibility of the social-media in "Web 3.0" — and a basic human right.

The Fate of the Species: Why the Human Race May Cause Its Own Extinction and How We Can Stop It
Fred Guterl BLOOMSBURY *224 pp. $25 (2012)*
Scientific American's executive editor, Fred Guterl, pulls no punches in this succinct round-up of the global trends that threaten humanity. He considers, in turn and backed by intriguing research, the rise of superviruses, rapid species extinctions, climate change, the disruption of ecosystems, synthetic biology and bioweaponry, and our over-dependence on machines. Ultimately, argues Guterl, the solutions lie in the very technology that propelled us into the current chaos — along with plain human adaptability.



Cracking the Egyptian Code: The Revolutionary Life of Jean-François Champollion
Andrew Robinson THAMES & HUDSON *272 pp. £19.95 (2012)*
In the first English-language biography of nineteenth-century "father of Egyptology" Jean-François Champollion, Andrew Robinson offers a vivid portrait of a prodigy, and richly contextualizes Champollion's work decoding the hieroglyphs on the Rosetta Stone. The book takes in Egyptomania from ancient Greece to eighteenth-century Britain, Champollion's rapid rise to professorhood, his rivalry with English polymath Thomas Young, years of preliminary work and travels in Egypt, and the advances that followed him — all beautifully illustrated.

Silent Spring Revisited
Conor Mark Jameson BLOOMSBURY *288 pp. £16.95 (2012)*
Natural-history writer Conor Jameson uses Rachel Carson's 1962 work *Silent Spring* as a focus for reflection on conservation and environmentalism in the decades since then. He begins with tens of thousands of UK birds dying in the 1960s, felled by pesticides, and moves through oil spills, the work of the UK Royal Society for the Protection of Birds and the steady decline in avian species. The 'silencing of spring', Jameson notes, continues — but reintroduction programmes, given the right support, are beacons in the gloom.

T. CHAPMAN

Q&A Bernie Krause

Soundscape explorer

Bioacoustician Bernie Krause has travelled the world for decades to gather animal sounds for his Wild Sanctuary archive (www.wildsanctuary.com). Following the release of his book about this work, The Great Animal Orchestra, *he talks about the calls of the wild.*

How did you first get into sound?
There were open fields where I grew up near Detroit, Michigan, and I remember the summer-evening sounds of insects and birds. But as with sex in the 1940s, there was no way to discuss it. In my teens I discovered the guitar, and in 1963, three years after graduating from university, I took Pete Seeger's slot in the band The Weavers. Then I moved to California, got into synthesizers and, with my late music partner Paul Beaver, contributed to more than 100 feature films, including *Apocalypse Now* (1979), *Invasion of the Body Snatchers* and *Doctor Doolittle* (1967).

Tell me about your 1970 album *In a Wild Sanctuary*.
This collaboration with Paul was the first recorded music to use long segments of wild sound. At the time, people would use a parabolic dish to isolate the sounds of bird species. I found stereo recordings more engaging. I went to the woods with a tape recorder and two microphones. I put on my headphones and the space opened up, calming me in a way that music didn't. I decided to record natural soundscapes for the rest of my life.

What is a soundscape?
Composer and naturalist R. Murray Schafer defined it as everything that reaches our ears in a given moment. There are three kinds of sound: geophony (from wind, rain, earthquakes and other natural, non-living sources); biophony (from non-human animals); and anthrophony (from humans and their machines). Anthrophony is getting harder to escape.

➲ NATURE.COM
For sound clips and an interview with Bernie Krause:
go.nature.com/5tlcza

The Great Animal Orchestra: Finding the Origins of Music in the World's Wild Places
BERNIE KRAUSE
Little, Brown/Profile Books: 2012. 288 pp. $26.99/£12.99

Do animals find acoustic niches?
In 1983 I was in an old-growth forest in Kenya recording for an exhibition at the California Academy of Sciences in San Francisco. Lying in my tent with headphones on, I suddenly heard hyenas, elephants, frogs and insects as an orchestrated collective, each singing within its own bandwidth. The creatures established both temporal and frequency niches for their vocalizations. I have found similar patterns in animal soundscapes around the world.

What can sound tell us about the health of an ecosystem?
When you record an unhealthy ecosystem or 'biome' — one that has been slashed and burned, for example — the voices tend to be faint and chaotic, like an untuned orchestra without a conductor and score. In 1988, I recorded in a site in the Sierra Nevada mountains before it was 'selectively logged' — a technique meant to have no impact on creature density and diversity of habitat. The place looked the same afterwards, but there was only sporadic birdsong, with almost no frogs or insects. Biomic elements find their niches over time, as I discovered in acoustic observations of older, more pristine habitats. I've recorded at that site 15 times over a number of years since the logging, and found that the biophony has not yet recovered.

How has technology changed your work?
When I started, I had to carry 80 kilograms of expensive equipment into the field to record for a month. Now I can record 10 times as much with less than 5 kilograms of gear. I can cover a site with large numbers of recording monitors and collect vast amounts of calibrated data for future reference. At the same time, we've lost an enormous amount of wild habitat worldwide, and human noise keeps encroaching on the places that remain. These days, to capture one hour of usable material, I must spend several hundred hours searching for undisturbed sites, avoiding human noise, or both. Half my archive is made up of soundscapes that no longer exist.

What are some of the most remarkable animal sounds you've heard?
Snapping shrimp stun their prey by creating cavitation bubbles with their claws; the bubble explodes like a starting pistol in your ear. I've caught the sound of red fire ants rubbing their hind legs on their abdomens to summon others to dig around a microphone and keep the entrance of their hole clear. One of the most frightening sounds I've recorded was of Ecuadorian baby vultures using a hollow tree to amplify their voices. I've also had close calls with a growling jaguar on an Amazon trail at night, and mountain gorillas in Rwanda, which emitted the loudest screams I've ever heard from any land animal before an attack.

Is it true that you have also recorded snow falling and maize (corn) growing?
Snow can't be recorded directly. But if you catch the right conditions, when it is almost freezing and the air is heavy with moisture, you can capture the sound by attaching a clip-on microphone to low-lying bushes. When the snow falls on the branches, it creates a little vibration. As for maize, one film I worked on sent me to Iowa to record it growing. I went out with microphones on a hot August night and waited. I discovered that maize grows each night by telescoping upwards. The stalks squeak like rubber balloons. ■

INTERVIEW BY JASCHA HOFFMAN

Correspondence

Don't let furore over neutrinos blur results

Neutrinos have been in the news again — and not just because of the debate over last year's OPERA experiment at the Gran Sasso National Laboratory in Italy, the results of which gave rise to the mistaken claim that the particles could travel faster than light (*Nature* **484,** 287–288; 2012). In March, multinational experiments at the Daya Bay reactor in Guangdong Province, China, tracked down a fundamental parameter that describes neutrino oscillations, and Fermilab's MINERvA experiment near Chicago, Illinois, transmitted a message using neutrino communication for the first time.

In the wake of these exciting findings, neutrino physicists should not be too tempted to release new data ahead of thorough analysis and ratification, which could harm science. Neutrinos are elusive and neutrino experiments are extraordinarily complex, so physicists should remain sceptical and ensure that results are solid before they are proclaimed to the tax-paying public.

Tommy Ohlsson *KTH Royal Institute of Technology, AlbaNova University Center, Stockholm, Sweden.* *tommy@theophys.kth.se*

Control electronic waste in India

Legislation that came into effect in India this month aims to deal with the environmental effects of electronic waste in the country. According to a government report, this waste stream has increased by a factor of more than five in seven years and is expected to exceed 800,000 tonnes in 2012.

However, the new law does not ban the dumping of toxic electronic waste from overseas, which contributes a further 50,000 tonnes. This is in violation of the 1992 Basel Convention, which restricts disposal and transboundary movements of hazardous waste, particularly from developed to developing countries.

India should follow China's example and stand firm against the dumping of electronic waste by the European Union and the United States, for example. It must tighten up enforcement of the Indian Supreme Court's 1997 blanket ban on the import and export of hazardous waste, in line with the Basel Convention.

To tackle India's domestic electronic waste, the new law stipulates that manufacturers must follow strict collection and recycling procedures, including a buy-back system (see go.nature.com/48fdyn). It is essential that these measures are rapidly implemented and then properly enforced by the state pollution-control boards.

Govindasamy Agoramoorthy *Tajen University, Yanpu, Pingtung, Taiwan.*
agoram@mail.tajen.edu.tw
Chiranjib Chakraborty *Vellore Institute of Technology (VIT) University, Tamil Nadu, India.*

Preserve Brazil's aquatic biodiversity

Brazil's aquatic biodiversity is under threat from a proposed law that aims to boost degraded fishery resources. If approved, the law — put forward by Nelson Meurer of the Brazilian National Congress — would allow the cultivation of non-native fish species in freshwater aquaculture cages, overriding the currently prohibited introduction of non-native species into Brazil.

The fish that would be introduced are tilapia and carp species, and other species that are potentially invasive in Brazil (J. R. S. Vitule *et al. Fish Fish.* **10,** 98–108; 2009). If these were to escape, they would further disrupt native freshwater biodiversity, which is already compromised by dam construction and pollution.

Politicians should instead be creating mechanisms to preserve native fauna and ecosystem functions, helping to realize Brazil's potential as a model for biodiversity conservation in the spirit of next month's Rio+20 conference. Meeting socio-economic needs must have the backing of sound environmental research.

Jean R. S. Vitule* *Federal University of Paraná, Curitiba, Paraná, Brazil.*
biovitule@gmail.com
**On behalf of 5 co-authors (for a full list, see go.nature.com/9yvvk1).*

In defence of the animal model

Jocelyn Rice points out perceived shortcomings of the experimental autoimmune encephalomyelitis mouse in modelling multiple sclerosis and in advancing effective human treatments for this disease (*Nature* **484,** S9; 2012; online only). However, her title ('Animal models: Not close enough') seems to cast doubt on the value of animal models in general for developing therapeutic strategies. Even if unintended, this implication undermines efforts to narrow the gap between research and the clinic.

We all share the author's frustration over the lag that separates direct medical benefits from animal research. But to shorten it, we should foster the possibilities of each model, not malign them.

Selective reporting of animal models that fail to deliver anticipated therapies risks promoting misperception among clinical researchers and policy-makers. They are more likely to remember that one mouse model was "worryingly unreliable" for screening multiple sclerosis treatments than they are to recall that mice were crucial in the development of many life-saving therapies.

Examples include ipilimumab, an antibody therapy that extends life in patients with metastatic melanoma; losartan, which reduces aortic disease in patients with Marfan syndrome; and the retinoic acid/arsenic trioxide therapy that saves the lives of patients with acute promyelocytic leukaemia.

Richard M. Baldarelli *The Jackson Laboratory, Bar Harbor, Maine, USA.*
richard.baldarelli@jax.org

The social sciences are already relevant

Luk Van Langenhove argues that the social sciences should be made more relevant (*Nature* **484,** 442; 2012). But the problem is rather that society remains largely unaware of the thousands of social-science studies produced every year that are relevant to global challenges such as climate change. Efforts should focus on increasing the societal use of scientific knowledge instead of producing more, starting with better communication of research findings to the public.

Van Langenhove notes that just 1,600 papers out of all social-science publications in 2010 (1.6%) contain the keywords 'environmental change' or 'climate change'. But keywords are not always the best way to rate the impact of different topics. Articles that might be relevant do not always mention the right keywords, and findings depend on keyword choice (for example, when I included 'sustainability' in Van Langenhove's search, a further 1,493 papers appeared). Social-sciences papers still score more hits for these keywords than papers in the natural sciences (0.65%).

Other topics relevant to society include cancer, AIDS and obesity, which I found to score 4,173, 3,341 and 2,365 social-science articles, respectively, for 2010. The social sciences are therefore already contributing substantially to solving societal issues.

Frank J. van Rijnsoever *Copernicus Institute of Sustainable Development, Utrecht University, the Netherlands.*
f.j.vanrijnsoever@uu.nl

Correspondence

Don't let furore over neutrinos blur results

Neutrinos have been in the news again — and not just because of the debate over last year's OPERA experiment at the Gran Sasso National Laboratory in Italy, the results of which gave rise to the mistaken claim that the particles could travel faster than light (*Nature* **484,** 287–288; 2012). In March, multinational experiments at the Daya Bay reactor in Guangdong Province, China, tracked down a fundamental parameter that describes neutrino oscillations, and Fermilab's MINERvA experiment near Chicago, Illinois, transmitted a message using neutrino communication for the first time.

In the wake of these exciting findings, neutrino physicists should not be too tempted to release new data ahead of thorough analysis and ratification, which could harm science. Neutrinos are elusive and neutrino experiments are extraordinarily complex, so physicists should remain sceptical and ensure that results are solid before they are proclaimed to the tax-paying public.

Tommy Ohlsson *KTH Royal Institute of Technology, AlbaNova University Center, Stockholm, Sweden.* *tommy@theophys.kth.se*

Control electronic waste in India

Legislation that came into effect in India this month aims to deal with the environmental effects of electronic waste in the country. According to a government report, this waste stream has increased by a factor of more than five in seven years and is expected to exceed 800,000 tonnes in 2012.

However, the new law does not ban the dumping of toxic electronic waste from overseas, which contributes a further 50,000 tonnes. This is in violation of the 1992 Basel Convention, which restricts disposal and transboundary movements of hazardous waste, particularly from developed to developing countries.

India should follow China's example and stand firm against the dumping of electronic waste by the European Union and the United States, for example. It must tighten up enforcement of the Indian Supreme Court's 1997 blanket ban on the import and export of hazardous waste, in line with the Basel Convention.

To tackle India's domestic electronic waste, the new law stipulates that manufacturers must follow strict collection and recycling procedures, including a buy-back system (see go.nature.com/48fdyn). It is essential that these measures are rapidly implemented and then properly enforced by the state pollution-control boards.

Govindasamy Agoramoorthy *Tajen University, Yanpu, Pingtung, Taiwan.* *agoram@mail.tajen.edu.tw*
Chiranjib Chakraborty *Vellore Institute of Technology (VIT) University, Tamil Nadu, India.*

Preserve Brazil's aquatic biodiversity

Brazil's aquatic biodiversity is under threat from a proposed law that aims to boost degraded fishery resources. If approved, the law — put forward by Nelson Meurer of the Brazilian National Congress — would allow the cultivation of non-native fish species in freshwater aquaculture cages, overriding the currently prohibited introduction of non-native species into Brazil.

The fish that would be introduced are tilapia and carp species, and other species that are potentially invasive in Brazil (J. R. S. Vitule *et al. Fish Fish.* **10,** 98–108; 2009). If these were to escape, they would further disrupt native freshwater biodiversity, which is already compromised by dam construction and pollution.

Politicians should instead be creating mechanisms to preserve native fauna and ecosystem functions, helping to realize Brazil's potential as a model for biodiversity conservation in the spirit of next month's Rio+20 conference. Meeting socio-economic needs must have the backing of sound environmental research.

Jean R. S. Vitule* *Federal University of Paraná, Curitiba, Paraná, Brazil.* *biovitule@gmail.com*
**On behalf of 5 co-authors (for a full list, see go.nature.com/9yvvk1).*

In defence of the animal model

Jocelyn Rice points out perceived shortcomings of the experimental autoimmune encephalomyelitis mouse in modelling multiple sclerosis and in advancing effective human treatments for this disease (*Nature* **484,** S9; 2012; online only). However, her title ('Animal models: Not close enough') seems to cast doubt on the value of animal models in general for developing therapeutic strategies. Even if unintended, this implication undermines efforts to narrow the gap between research and the clinic.

We all share the author's frustration over the lag that separates direct medical benefits from animal research. But to shorten it, we should foster the possibilities of each model, not malign them.

Selective reporting of animal models that fail to deliver anticipated therapies risks promoting misperception among clinical researchers and policy-makers. They are more likely to remember that one mouse model was "worryingly unreliable" for screening multiple sclerosis treatments than they are to recall that mice were crucial in the development of many life-saving therapies.

Examples include ipilimumab, an antibody therapy that extends life in patients with metastatic melanoma; losartan, which reduces aortic disease in patients with Marfan syndrome; and the retinoic acid/arsenic trioxide therapy that saves the lives of patients with acute promyelocytic leukaemia.

Richard M. Baldarelli *The Jackson Laboratory, Bar Harbor, Maine, USA.* *richard.baldarelli@jax.org*

The social sciences are already relevant

Luk Van Langenhove argues that the social sciences should be made more relevant (*Nature* **484,** 442; 2012). But the problem is rather that society remains largely unaware of the thousands of social-science studies produced every year that are relevant to global challenges such as climate change. Efforts should focus on increasing the societal use of scientific knowledge instead of producing more, starting with better communication of research findings to the public.

Van Langenhove notes that just 1,600 papers out of all social-science publications in 2010 (1.6%) contain the keywords 'environmental change' or 'climate change'. But keywords are not always the best way to rate the impact of different topics. Articles that might be relevant do not always mention the right keywords, and findings depend on keyword choice (for example, when I included 'sustainability' in Van Langenhove's search, a further 1,493 papers appeared). Social-sciences papers still score more hits for these keywords than papers in the natural sciences (0.65%).

Other topics relevant to society include cancer, AIDS and obesity, which I found to score 4,173, 3,341 and 2,365 social-science articles, respectively, for 2010. The social sciences are therefore already contributing substantially to solving societal issues.

Frank J. van Rijnsoever *Copernicus Institute of Sustainable Development, Utrecht University, the Netherlands.* *f.j.vanrijnsoever@uu.nl*

Correspondence

Don't let furore over neutrinos blur results

Neutrinos have been in the news again — and not just because of the debate over last year's OPERA experiment at the Gran Sasso National Laboratory in Italy, the results of which gave rise to the mistaken claim that the particles could travel faster than light (*Nature* **484,** 287–288; 2012). In March, multinational experiments at the Daya Bay reactor in Guangdong Province, China, tracked down a fundamental parameter that describes neutrino oscillations, and Fermilab's MINERvA experiment near Chicago, Illinois, transmitted a message using neutrino communication for the first time.

In the wake of these exciting findings, neutrino physicists should not be too tempted to release new data ahead of thorough analysis and ratification, which could harm science. Neutrinos are elusive and neutrino experiments are extraordinarily complex, so physicists should remain sceptical and ensure that results are solid before they are proclaimed to the tax-paying public.

Tommy Ohlsson *KTH Royal Institute of Technology, AlbaNova University Center, Stockholm, Sweden.* *tommy@theophys.kth.se*

Control electronic waste in India

Legislation that came into effect in India this month aims to deal with the environmental effects of electronic waste in the country. According to a government report, this waste stream has increased by a factor of more than five in seven years and is expected to exceed 800,000 tonnes in 2012.

However, the new law does not ban the dumping of toxic electronic waste from overseas, which contributes a further 50,000 tonnes. This is in violation of the 1992 Basel Convention, which restricts disposal and transboundary movements of hazardous waste, particularly from developed to developing countries.

India should follow China's example and stand firm against the dumping of electronic waste by the European Union and the United States, for example. It must tighten up enforcement of the Indian Supreme Court's 1997 blanket ban on the import and export of hazardous waste, in line with the Basel Convention.

To tackle India's domestic electronic waste, the new law stipulates that manufacturers must follow strict collection and recycling procedures, including a buy-back system (see go.nature.com/48fdyn). It is essential that these measures are rapidly implemented and then properly enforced by the state pollution-control boards.

Govindasamy Agoramoorthy *Tajen University, Yanpu, Pingtung, Taiwan.* *agoram@mail.tajen.edu.tw*
Chiranjib Chakraborty *Vellore Institute of Technology (VIT) University, Tamil Nadu, India.*

Preserve Brazil's aquatic biodiversity

Brazil's aquatic biodiversity is under threat from a proposed law that aims to boost degraded fishery resources. If approved, the law — put forward by Nelson Meurer of the Brazilian National Congress — would allow the cultivation of non-native fish species in freshwater aquaculture cages, overriding the currently prohibited introduction of non-native species into Brazil.

The fish that would be introduced are tilapia and carp species, and other species that are potentially invasive in Brazil (J. R. S. Vitule *et al. Fish Fish.* **10,** 98–108; 2009). If these were to escape, they would further disrupt native freshwater biodiversity, which is already compromised by dam construction and pollution.

Politicians should instead be creating mechanisms to preserve native fauna and ecosystem functions, helping to realize Brazil's potential as a model for biodiversity conservation in the spirit of next month's Rio+20 conference. Meeting socio-economic needs must have the backing of sound environmental research.

Jean R. S. Vitule* *Federal University of Paraná, Curitiba, Paraná, Brazil.* *biovitule@gmail.com*
**On behalf of 5 co-authors (for a full list, see go.nature.com/9yvvk1).*

In defence of the animal model

Jocelyn Rice points out perceived shortcomings of the experimental autoimmune encephalomyelitis mouse in modelling multiple sclerosis and in advancing effective human treatments for this disease (*Nature* **484,** S9; 2012; online only). However, her title ('Animal models: Not close enough') seems to cast doubt on the value of animal models in general for developing therapeutic strategies. Even if unintended, this implication undermines efforts to narrow the gap between research and the clinic.

We all share the author's frustration over the lag that separates direct medical benefits from animal research. But to shorten it, we should foster the possibilities of each model, not malign them.

Selective reporting of animal models that fail to deliver anticipated therapies risks promoting misperception among clinical researchers and policy-makers. They are more likely to remember that one mouse model was "worryingly unreliable" for screening multiple sclerosis treatments than they are to recall that mice were crucial in the development of many life-saving therapies.

Examples include ipilimumab, an antibody therapy that extends life in patients with metastatic melanoma; losartan, which reduces aortic disease in patients with Marfan syndrome; and the retinoic acid/arsenic trioxide therapy that saves the lives of patients with acute promyelocytic leukaemia.

Richard M. Baldarelli *The Jackson Laboratory, Bar Harbor, Maine, USA.* *richard.baldarelli@jax.org*

The social sciences are already relevant

Luk Van Langenhove argues that the social sciences should be made more relevant (*Nature* **484,** 442; 2012). But the problem is rather that society remains largely unaware of the thousands of social-science studies produced every year that are relevant to global challenges such as climate change. Efforts should focus on increasing the societal use of scientific knowledge instead of producing more, starting with better communication of research findings to the public.

Van Langenhove notes that just 1,600 papers out of all social-science publications in 2010 (1.6%) contain the keywords 'environmental change' or 'climate change'. But keywords are not always the best way to rate the impact of different topics. Articles that might be relevant do not always mention the right keywords, and findings depend on keyword choice (for example, when I included 'sustainability' in Van Langenhove's search, a further 1,493 papers appeared). Social-sciences papers still score more hits for these keywords than papers in the natural sciences (0.65%).

Other topics relevant to society include cancer, AIDS and obesity, which I found to score 4,173, 3,341 and 2,365 social-science articles, respectively, for 2010. The social sciences are therefore already contributing substantially to solving societal issues.

Frank J. van Rijnsoever *Copernicus Institute of Sustainable Development, Utrecht University, the Netherlands.* *f.j.vanrijnsoever@uu.nl*

Correspondence

Don't let furore over neutrinos blur results

Neutrinos have been in the news again — and not just because of the debate over last year's OPERA experiment at the Gran Sasso National Laboratory in Italy, the results of which gave rise to the mistaken claim that the particles could travel faster than light (*Nature* **484,** 287–288; 2012). In March, multinational experiments at the Daya Bay reactor in Guangdong Province, China, tracked down a fundamental parameter that describes neutrino oscillations, and Fermilab's MINERvA experiment near Chicago, Illinois, transmitted a message using neutrino communication for the first time.

In the wake of these exciting findings, neutrino physicists should not be too tempted to release new data ahead of thorough analysis and ratification, which could harm science. Neutrinos are elusive and neutrino experiments are extraordinarily complex, so physicists should remain sceptical and ensure that results are solid before they are proclaimed to the tax-paying public.

Tommy Ohlsson *KTH Royal Institute of Technology, AlbaNova University Center, Stockholm, Sweden.* *tommy@theophys.kth.se*

Control electronic waste in India

Legislation that came into effect in India this month aims to deal with the environmental effects of electronic waste in the country. According to a government report, this waste stream has increased by a factor of more than five in seven years and is expected to exceed 800,000 tonnes in 2012.

However, the new law does not ban the dumping of toxic electronic waste from overseas, which contributes a further 50,000 tonnes. This is in violation of the 1992 Basel Convention, which restricts disposal and transboundary movements of hazardous waste, particularly from developed to developing countries.

India should follow China's example and stand firm against the dumping of electronic waste by the European Union and the United States, for example. It must tighten up enforcement of the Indian Supreme Court's 1997 blanket ban on the import and export of hazardous waste, in line with the Basel Convention.

To tackle India's domestic electronic waste, the new law stipulates that manufacturers must follow strict collection and recycling procedures, including a buy-back system (see go.nature.com/48fdyn). It is essential that these measures are rapidly implemented and then properly enforced by the state pollution-control boards.

Govindasamy Agoramoorthy *Tajen University, Yanpu, Pingtung, Taiwan.*
agoram@mail.tajen.edu.tw
Chiranjib Chakraborty *Vellore Institute of Technology (VIT) University, Tamil Nadu, India.*

Preserve Brazil's aquatic biodiversity

Brazil's aquatic biodiversity is under threat from a proposed law that aims to boost degraded fishery resources. If approved, the law — put forward by Nelson Meurer of the Brazilian National Congress — would allow the cultivation of non-native fish species in freshwater aquaculture cages, overriding the currently prohibited introduction of non-native species into Brazil.

The fish that would be introduced are tilapia and carp species, and other species that are potentially invasive in Brazil (J. R. S. Vitule *et al. Fish Fish.* **10,** 98–108; 2009). If these were to escape, they would further disrupt native freshwater biodiversity, which is already compromised by dam construction and pollution.

Politicians should instead be creating mechanisms to preserve native fauna and ecosystem functions, helping to realize Brazil's potential as a model for biodiversity conservation in the spirit of next month's Rio+20 conference. Meeting socio-economic needs must have the backing of sound environmental research.

Jean R. S. Vitule* *Federal University of Paraná, Curitiba, Paraná, Brazil.*
biovitule@gmail.com
**On behalf of 5 co-authors (for a full list, see go.nature.com/9yvvk1).*

In defence of the animal model

Jocelyn Rice points out perceived shortcomings of the experimental autoimmune encephalomyelitis mouse in modelling multiple sclerosis and in advancing effective human treatments for this disease (*Nature* **484,** S9; 2012; online only). However, her title ('Animal models: Not close enough') seems to cast doubt on the value of animal models in general for developing therapeutic strategies. Even if unintended, this implication undermines efforts to narrow the gap between research and the clinic.

We all share the author's frustration over the lag that separates direct medical benefits from animal research. But to shorten it, we should foster the possibilities of each model, not malign them.

Selective reporting of animal models that fail to deliver anticipated therapies risks promoting misperception among clinical researchers and policy-makers. They are more likely to remember that one mouse model was "worryingly unreliable" for screening multiple sclerosis treatments than they are to recall that mice were crucial in the development of many life-saving therapies.

Examples include ipilimumab, an antibody therapy that extends life in patients with metastatic melanoma; losartan, which reduces aortic disease in patients with Marfan syndrome; and the retinoic acid/arsenic trioxide therapy that saves the lives of patients with acute promyelocytic leukaemia.

Richard M. Baldarelli *The Jackson Laboratory, Bar Harbor, Maine, USA.*
richard.baldarelli@jax.org

The social sciences are already relevant

Luk Van Langenhove argues that the social sciences should be made more relevant (*Nature* **484,** 442; 2012). But the problem is rather that society remains largely unaware of the thousands of social-science studies produced every year that are relevant to global challenges such as climate change. Efforts should focus on increasing the societal use of scientific knowledge instead of producing more, starting with better communication of research findings to the public.

Van Langenhove notes that just 1,600 papers out of all social-science publications in 2010 (1.6%) contain the keywords 'environmental change' or 'climate change'. But keywords are not always the best way to rate the impact of different topics. Articles that might be relevant do not always mention the right keywords, and findings depend on keyword choice (for example, when I included 'sustainability' in Van Langenhove's search, a further 1,493 papers appeared). Social-sciences papers still score more hits for these keywords than papers in the natural sciences (0.65%).

Other topics relevant to society include cancer, AIDS and obesity, which I found to score 4,173, 3,341 and 2,365 social-science articles, respectively, for 2010. The social sciences are therefore already contributing substantially to solving societal issues.

Frank J. van Rijnsoever *Copernicus Institute of Sustainable Development, Utrecht University, the Netherlands.*
f.j.vanrijnsoever@uu.nl

TURNING POINT Director of sustainability institute is working to curb pollution **p.407**

CAREERS BLOG The latest discussions and news on research jobs **go.nature.com/ielkkf**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

A.RADOSAVLJEVIC

PUBLISHING

Going digital

Creating electronic textbooks requires ingenuity, teamwork and multimedia savvy.

BY ROBERTA KWOK

Douglas Emlen is hard at work on an evolution textbook. But this is not just a print book. Creating an iPad app with images, audio and video clips, and interactive graphics and exercises has meant collaborating with designers, programmers and an artist on a digital version of the book.

Emlen, an evolutionary biologist at the University of Montana in Missoula — who is co-writing the book with Carl Zimmer, a science writer based in Guilford, Connecticut — is part of an emerging group of scientists navigating the world of digital textbooks. The idea of electronic instructional materials is not new: texts in e-book form, as well as online supplements, teaching tools and homework systems have been available for years. But as tablets and e-books become more popular, publishers are increasingly placing equal or greater importance on the digital product rather than considering it as an add-on to the printed book. Some publishers are moving towards electronic-only textbooks. A survey released in March by the Pearson Foundation in Mill Valley, California, which promotes literacy and education, showed that the proportion of university students who own tablets grew from 7% in 2011 to 25% in 2012. More than two-thirds of university students have used a digital textbook, the survey says, and more than half prefer the digital format to print. The increasing popularity means that authors must consider the digital vision of the book when coming up with an idea and work with diverse teams to weave together text, multimedia and interactive exercises and quizzes.

"The role of an author in the past was, 'Let me write a big manuscript and mail it in to you,'" says Kurt Strand, senior vice-president and chief product officer at McGraw-Hill Higher Education in Dubuque, Iowa. Now, the author provides the vision for the complete learning experience, he says.

Authors must consider the most effective use of multimedia, adapt to the changing structure of textbooks and be flexible in response to feedback from user-testing. Although the financial reward of such projects remains uncertain, some digital-textbook authors have found satisfaction in exploring alternative ways to teach concepts and potentially improve their connections with students.

The first electronic textbooks were little more than replicas of the print versions. But, with the release of the iPad, greater Internet bandwidth in schools and a growing popularity with students, textbooks with more interactive features are emerging. In January, Apple announced the release of its iBooks 2, an app with improved support for digital textbooks, and announced that three major publishers would sell their textbooks through its online shop. Inkling, a start-up company in San Francisco, California, now has more than 100 digital textbooks available through its iPad app and will soon make them available through web browsers. And in February, the US Federal Communications Commission in Washington DC urged schools and companies to supply all primary and secondary education students in the United States with at least one digital textbook within five years.

The shift from print to digital textbooks is happening very quickly, says Morgan Ryan, project director of *E.O. Wilson's Life on Earth*, a digital-only biology textbook being developed by the E.O. Wilson Biodiversity Foundation in Chapel Hill, North Carolina. "There's this feeling that it's finally arrived," he says.

Sceptics warn that students may become lost among, or distracted by, some of the ▶

electronic bells and whistles. Nevertheless, publishers are taking the plunge, and they are expecting authors to have an active role in shaping the product.

A DIGITAL VISION

Even as textbooks shift to a different medium, publishers say that authors still need to understand the challenge that they have taken on, and to have an effective, classroom-tested approach to teaching. "I have had authors come forward and say, 'My distinguishing feature will be that this book will be digital'," says Kaye Pace, a vice-president and executive publisher in the global education group at John Wiley & Sons in Hoboken, New Jersey. "I don't think that works. You have to start with, 'What is the issue that you're trying to resolve?'"

Proposals to publishers should include ideas for multimedia and interactivity. Although publishers generally don't expect authors to provide app prototypes or refined illustrations, they do want specific concepts that can be executed by the publishing team. Storyboards are often sufficient. Authors may not ever learn all the ins and outs of three-dimensional graphics and programming, but they do have to think about how the medium is going to be used, notes Ryan.

Some authors are planning graphics and interactive elements up front, instead of writing the text first and deciding which multimedia to add later. "The digital product becomes much more of a teaching tool than a way of illustrating in some visual form what the words are saying," says Eric Schulz, a mathematics instructor at Walla Walla Community College in Washington, who created roughly 650 interactive figures for a calculus textbook (published in 2010 by Pearson).

INKLING

Inkling founder Matt MacInnis says authors need to think of new ways to communicate with learners.

But authors also need to be aware that multimedia and interactive elements are expensive to produce. The amount of money available will depend, in part, on the size of the potential market: an introductory economics textbook, for example, will probably have a larger budget than a niche upper-level textbook about community ecology. Nonetheless, some publishers encourage authors to start with a grand vision. "I'd rather start there than get something that's less exciting," says Ben Roberts from Roberts and Company Publishers in Greenwood Village, Colorado, which is publishing Emlen and Zimmer's book. Elements can be discarded later if need be.

RECONCEIVING THE TEXTBOOK

Cost and time considerations are not the only reasons to use multimedia judiciously. "Students aren't going to learn more just because you throw a whole bunch of videos in there," says Emlen. Instead, authors should carefully consider which method would be most appropriate for achieving their instructional goal. For example, David Johnston, a marine biologist at Duke University Marine Laboratory in Beaufort, North Carolina, and project leader of *Cachalot*, a self-published digital textbook about marine megafauna (see 'Self-publishing'), says that it makes sense to use a picture to show the features of a penguin's tongue and an audio clip to demonstrate the noises in echolocation, but to illustrate the rate of sea-ice decline, a two-dimensional graph could suffice. And complex material, such as a set of equations, may be difficult to learn if it is presented only in a transient format, such as animation.

Opinions differ on how large a role the text should have compared with graphics, animations and interactive features. "I often see too heavy a reliance on expository text," says Matt MacInnis, founder and chief executive of Inkling, noting that Inkling's data suggest that learners skim, search and refer to text instead of reading it. "Find ways to be brief and multimodal rather than expository and textual." For example, the next iteration of WileyPLUS, Wiley's online teaching and

SELF-PUBLISHING

A time-intensive project that may not be for everyone

With the release of Apple's iBooks Author e-book authoring app and the ability to distribute mobile apps easily, obstacles to self-publishing a textbook continue to fall. But is it a good idea? It may be, for those who have the time to invest in an intensive project and to oversee everything from art to marketing.

Self-publishing proved to be a satisfying option for David Johnston, a marine biologist at the Duke University Marine Laboratory in Beaufort, North Carolina. Publishers turned down Johnston's proposal for a digital textbook about marine megafauna. So he got a grant from the Duke Center for Instructional Technology, recruited computer science students to write an app, solicited text from about 30 experts and assembled multimedia from scientists, National Geographic's Crittercam and the Woods Hole Oceanographic Institution. Within a year, his team had released an iPad app, *Cachalot*, which offers multimedia-enriched entries about marine animals, open-access articles and teaching modules. "If you're a creative individual and you have time, then it's a great way for you to get across the ideas that you want," says Johnston.

Cachalot has been downloaded more than 3,500 times, and three universities have expressed interest in using it. The app is free, so the project has certainly not made Johnston rich, but he says it has allowed him to connect with other experts in the field through their contribution to the project, brought him recognition within his department and prompted the marine lab to consider buying iPads for students to use.

But self-publishing authors take on a heavy burden. They must fill in services that publishers would provide, such as editing, art, design, programming, marketing, sales and customer service, says Morgan Ryan, project director of the digital textbook *E.O. Wilson's Life on Earth,* which is being developed by the E.O. Wilson Biodiversity Foundation in Chapel Hill, North Carolina. The amount of time and money needed depends on the scope of the project. An author who writes the text and uses free multimedia could publish a textbook using iBooks Author essentially for free. Johnston, who focuses on a very specific topic and gets help from students and colleagues, says that he spends about five hours a week on *Cachalot* and so far has been awarded US$15,000 in grants. Hiring professional contractors could run into the tens or hundreds of thousands of dollars.

The decision may boil down to whether an author is willing to be a manager and publisher. "You're going to have to put on a hat that you're probably not used to," says online-teaching tool developer Dan Johnson, a senior biology lecturer at Wake Forest University in Winston-Salem, North Carolina. **R.K.**

learning system, will 'chunk' material into concept modules and avoid long passages of unbroken text. But that doesn't mean all writing must be condensed into Twitter-sized bites. "It's going to come down to the writing, regardless of whether it's in electronic format or on paper," says Emlen.

The structure of textbooks is also in flux. Digital textbooks are becoming increasingly modular, as many publishers are selling individual chapters and allowing teachers to build customized versions. Some textbook producers are also migrating towards more open-ended navigation in which students can skip to topics rather than follow the linear ordering used in print. "Think about it as building a big website as opposed to building a book," says MacInnis. This approach may not work for every subject; trigonometry, for example, requires some linear progression.

Once a publisher has accepted the concept, the author needs to guide the publishing team that will execute the idea. Authors have to be at the centre of the creative process, says Roberts. "There will be animators, developers and instructional designers all trying to get their calling orders from the author." Authors must clearly communicate their vision and be prepared to iterate it as elements such as artwork and simulations are developed. And, because publishers are still working on the best approach to digital textbooks, authors also need to adjust their vision in response to feedback from testing with potential users. Compared with print textbooks, the authors have to be much more flexible and more attentive to the learner, says Strand.

Aspiring authors should use existing science apps and digital textbooks. They can get their feet wet by being a reviewer for one or authoring a component of a digital textbook or online learning system, such as a module, simulation, case study or assessment questions. For example, about 40 authors and reviewers worked on the digital *Principles of Biology* textbook from Nature Education, a division of Nature Publishing Group, which publishes *Nature*. And authors should consider teaming up with a multimedia-savvy partner, especially if the person is also an expert in their subject area. Graduate students may already have those skills, says Ryan.

In the end, authors should not get distracted from the core task of trying and validating better teaching methods. What makes a great author has not changed, says Susan Winslow, a publisher for life sciences at WH Freeman in New York (WH Freeman is owned by Macmillan, which also publishes *Nature*). "They're capable, in any medium, of connecting the dots for a novice." ■

Roberta Kwok *is a freelance writer in Burlingame, California.*

TURNING POINT
Mark Lawrence

Atmospheric scientist Mark Lawrence was named scientific director of the Institute for Advanced Sustainability Studies in Potsdam, Germany, last October, after a 19-year research stint at the Max Planck Institute for Chemistry in Mainz.

What drew you to Earth and atmospheric sciences?
I intended to study medicine at university, but after learning about the realities of being a doctor, I decided to move into basic physics. Towards the end of my undergraduate studies at the Georgia Institute of Technology in Atlanta, I spent a couple of semesters at the tyre manufacturer Michelin working on cutting edge car-suspension systems and ways to analyse how roads affect car tyres. However, I realized it was important to me that my work had a societal impact, so I switched to Earth and atmospheric sciences, which gave me the chance to combine broad aspects of physics as well as chemistry, and even some topics related to biology. It was an intellectually challenging science.

How did you end up in Germany and why have you stayed?
I wanted to see the world, and Germany seemed to be a good place because my now-wife was heading back to Germany to finish her studies. I had a US National Science Foundation graduate fellowship, so I contacted the then-director of the Max Planck Institute for Chemistry, atmospheric chemist Paul Crutzen, who agreed to supervise my thesis. I had planned to return to the United States after my PhD, but Paul jointly won the Nobel Prize in Chemistry in 1995 for his part in work on the ozone layer, and he persuaded me to stay at the institute for a postdoc, studying atmospheric models and the pollution outflow to the Indian Ocean — preparation for the 'Indian Ocean Experiment'. At the same time, qualifications started to take priority in German academia, and it became easier for young researchers to progress. In 2000, I won a grant from the German Federal Ministry of Education and Research to lead a junior research group — a five-year funding opportunity for young scientists seeking to establish themselves in the field — focused mainly on large-scale atmospheric pollution in the tropics, especially Asia.

How did you get your new post?
Through Paul, I met Klaus Töpfer, founding director of the Institute for Advanced Sustainability Studies. I gave him some of my climate papers to read, and he later invited me to Potsdam. The research collaboration grew from there. When the institute started looking for a scientific director, I was fortunate enough to be the right person in the right place at the right time.

My group is focusing on how humans are modifying the composition of the atmosphere and how this affects human life — for example; how ozone, methane and soot modify the composition of the atmosphere. We hope to find ways to make our modern lifestyles more sustainable by reducing pollution.

What is your secret for success?
I have a passion — I want to make a contribution to society through science. My job is more of a calling than a career. Honest self-analysis is important for professional success. We are very good at analysing our environment, but we should also look at ourselves and ask, 'What are our personal strengths and preferences, and where can we make a difference?'. Students should figure out what they can really do well that will make a positive contribution.

How do you juggle your work and personal life?
I spend weekends with my wife and children in Mainz, but my weekdays are concentrated on work. To focus, I usually get up early, meditate, go jogging and take a cold shower. Then I work until midnight. I find it is sustainable — the magic word — only if I keep my focus on my scientific contributions rather than on my career accomplishments, such as publication and citation numbers. ■

INTERVIEW BY ALEXANDRA BELL

learning system, will 'chunk' material into concept modules and avoid long passages of unbroken text. But that doesn't mean all writing must be condensed into Twitter-sized bites. "It's going to come down to the writing, regardless of whether it's in electronic format or on paper," says Emlen.

The structure of textbooks is also in flux. Digital textbooks are becoming increasingly modular, as many publishers are selling individual chapters and allowing teachers to build customized versions. Some textbook producers are also migrating towards more open-ended navigation in which students can skip to topics rather than follow the linear ordering used in print. "Think about it as building a big website as opposed to building a book," says MacInnis. This approach may not work for every subject; trigonometry, for example, requires some linear progression.

Once a publisher has accepted the concept, the author needs to guide the publishing team that will execute the idea. Authors have to be at the centre of the creative process, says Roberts. "There will be animators, developers and instructional designers all trying to get their calling orders from the author." Authors must clearly communicate their vision and be prepared to iterate it as elements such as artwork and simulations are developed. And, because publishers are still working on the best approach to digital textbooks, authors also need to adjust their vision in response to feedback from testing with potential users. Compared with print textbooks, the authors have to be much more flexible and more attentive to the learner, says Strand.

Aspiring authors should use existing science apps and digital textbooks. They can get their feet wet by being a reviewer for one or authoring a component of a digital textbook or online learning system, such as a module, simulation, case study or assessment questions. For example, about 40 authors and reviewers worked on the digital *Principles of Biology* textbook from Nature Education, a division of Nature Publishing Group, which publishes *Nature*. And authors should consider teaming up with a multimedia-savvy partner, especially if the person is also an expert in their subject area. Graduate students may already have those skills, says Ryan.

In the end, authors should not get distracted from the core task of trying and validating better teaching methods. What makes a great author has not changed, says Susan Winslow, a publisher for life sciences at WH Freeman in New York (WH Freeman is owned by Macmillan, which also publishes *Nature*). "They're capable, in any medium, of connecting the dots for a novice." ■

Roberta Kwok *is a freelance writer in Burlingame, California.*

TURNING POINT
Mark Lawrence

Atmospheric scientist Mark Lawrence was named scientific director of the Institute for Advanced Sustainability Studies in Potsdam, Germany, last October, after a 19-year research stint at the Max Planck Institute for Chemistry in Mainz.

What drew you to Earth and atmospheric sciences?

I intended to study medicine at university, but after learning about the realities of being a doctor, I decided to move into basic physics. Towards the end of my undergraduate studies at the Georgia Institute of Technology in Atlanta, I spent a couple of semesters at the tyre manufacturer Michelin working on cutting edge car-suspension systems and ways to analyse how roads affect car tyres. However, I realized it was important to me that my work had a societal impact, so I switched to Earth and atmospheric sciences, which gave me the chance to combine broad aspects of physics as well as chemistry, and even some topics related to biology. It was an intellectually challenging science.

How did you end up in Germany and why have you stayed?

I wanted to see the world, and Germany seemed to be a good place because my now-wife was heading back to Germany to finish her studies. I had a US National Science Foundation graduate fellowship, so I contacted the then-director of the Max Planck Institute for Chemistry, atmospheric chemist Paul Crutzen, who agreed to supervise my thesis. I had planned to return to the United States after my PhD, but Paul jointly won the Nobel Prize in Chemistry in 1995 for his part in work on the ozone layer, and he persuaded me to stay at the institute for a postdoc, studying atmospheric models and the pollution outflow to the Indian Ocean — preparation for the 'Indian Ocean Experiment'. At the same time, qualifications started to take priority in German academia, and it became easier for young researchers to progress. In 2000, I won a grant from the German Federal Ministry of Education and Research to lead a junior research group — a five-year funding opportunity for young scientists seeking to establish themselves in the field — focused mainly on large-scale atmospheric pollution in the tropics, especially Asia.

How did you get your new post?

Through Paul, I met Klaus Töpfer, founding director of the Institute for Advanced Sustainability Studies. I gave him some of my climate papers to read, and he later invited me to Potsdam. The research collaboration grew from there. When the institute started looking for a scientific director, I was fortunate enough to be the right person in the right place at the right time.

My group is focusing on how humans are modifying the composition of the atmosphere and how this affects human life — for example; how ozone, methane and soot modify the composition of the atmosphere. We hope to find ways to make our modern lifestyles more sustainable by reducing pollution.

What is your secret for success?

I have a passion — I want to make a contribution to society through science. My job is more of a calling than a career. Honest self-analysis is important for professional success. We are very good at analysing our environment, but we should also look at ourselves and ask, 'What are our personal strengths and preferences, and where can we make a difference?'. Students should figure out what they can really do well that will make a positive contribution.

How do you juggle your work and personal life?

I spend weekends with my wife and children in Mainz, but my weekdays are concentrated on work. To focus, I usually get up early, meditate, go jogging and take a cold shower. Then I work until midnight. I find it is sustainable — the magic word — only if I keep my focus on my scientific contributions rather than on my career accomplishments, such as publication and citation numbers. ■

INTERVIEW BY ALEXANDRA BELL

RAVAGES OF TIME

The generation gap.

BY ALEX SHVARTSMAN

JACEY

Jake Turner sat behind his desk, eyes closed, letting the muted sounds of music and laughter that emanated from the street wash over him. Outside, people were celebrating Ship Week.

The government had declared a planet-wide holiday and everyone was having a good time, except for a handful of unfortunate souls stuck in their jobs. After all, Ship Week happened only once every 45 years. Turner volunteered to work through the holiday, and his superiors, desperate for manpower, approved his request to cancel medical leave and come back to work.

A light chime announced his next appointment, forcing him out of his reverie. The medication he had taken earlier was beginning to wear off, and small pings of pain were tingling deep within his bones.

The door opened to admit a middle-aged woman dressed in a style that was decades out of fashion.

"Mrs Grobinski," he rose to greet her. "I'm Security Chief Turner."

Anna Grobinski smiled meekly and shook his hand.

"Please, sit." Turner stared at the off-worlder. "I understand that there was an incident involving your son during the previous Ship Week. He was left behind?"

The Ship shuttled between the inhabited star systems, delivering everything from medical advances to films, books, music and gossip from the other planets. It also carried migrants, people looking for a fresh start on another world. Those travelling to planets farther along the Ship's route welcomed the opportunity to spend a week exploring an exotic new world.

"It was an accident!" Her voice trembled. "By the time anyone realized Julek wasn't on one of the shuttles, it was too late. The Captain wouldn't delay departure …" She trailed off, her eyes filling with tears.

"I'm very sorry," Turner said, keeping his voice even. "I realize that it's only been a few months for you, since you lost him. But for your son, half a century has passed. He's older than you now."

Grobinski nodded, blotting under her eyes with a handkerchief.

"It says here," Turner pointed at his screen, "that you refused to disembark at Astor Prime and remained on board while the Ship made its rounds and returned to our star system. You understand about the time dilation, that your son would be 59 now, if he is still alive at all. So why have you come back?"

"I had to know," she said. "Can you imagine being completely alone when you're only 14? He must've been so scared. I had to know that he is all right. Know that he made a life for himself. That he forgives me." The tears were beginning to well up again.

Turner tensed up. This was going to be the hard part. The pain in his bones was really flaring up now. He welcomed it, a fitting punishment for what he had to do.

"That is why we're here," said Turner. "I've located Julek, and can assure you that he has done well for himself. He was adopted by a nice family, grew up and started a family of his own. You're a great-grandmother, Mrs Grobinski."

She exhaled, processing the news. "When can I see him?" she asked eagerly.

"I'm sorry to say that he chose not to meet with you in person," said Turner. "He felt that seeing him as an older man would be much too painful for you. Why, you probably wouldn't even recognize him."

Grobinski bit her lip, hard. "Do you have children, Mr Turner?"

He nodded.

"Then you should understand. You'd always recognize them, no matter the ravages of time. Always. There's a bond."

"I'm only passing along his wishes," said Turner. "He said you should continue on to Astor Prime. Make a new life for yourself there, like you always wanted to. It would make him happy to know that you'd moved on with your life."

Turner rose from his chair to indicate that their appointment was at an end. There were so many other things to do, so many issues to deal with, during Ship Week. Grobinski made no move to leave. She remained seated, staring at an undecorated wall with a forlorn expression on her face.

"Is there any message you'd like me to pass along," Turner prodded. "A letter, perhaps? Forward it to my office, and I'll make certain your son receives it."

"No letter," Grobinski finally said. "But … would you give him this?" She retrieved an antique pocket watch from her purse. "It was his father's. The lid was broken off and Julek always carried it with him, as a good luck charm."

"I'll pass it along," Turner promised.

When she was gone, Turner brought up the photos of his own family on his monitor. There was his wife, his children and grandchildren, his adopted parents and his younger stepbrother, who couldn't pronounce his name right as a three-year-old, and who was the first to begin calling him Jake.

Turner opened his desk drawer and took out the golden pocket-watch lid he kept there, next to the cancer pills. He pressed it to the watch and held the two pieces together for a long time, willing them to be whole. ■

Alex Shvartsman *is a writer and game designer. His adventures so far have included travelling to more than 30 countries, playing a card game for a living and building a successful business. He blogs at www.alexshvartsman.com.*

➲ NATURE.COM
Follow Futures on Facebook at:
go.nature.com/mtoodm

nature OUTLOOK

DIABETES

17 May 2012 / Vol 485 / Issue No. 7398

Cover art: Nik Spencer

The distinct but biologically related disorders that share the name diabetes impose vast human and economic losses. The US Center for Disease Control and Prevention estimates that medical expenses for people living with diabetes in the United States are, on average, 2.3 times higher than for non-diabetics. According to the World Health Organization, 346 million people worldwide — roughly the combined populations of the United States and Canada — have diabetes (page S2).

In the developing world, Type 2 diabetes is growing at an alarming rate as people gain access to the trappings of modernity — Western-style diets along with a more sedentary lifestyle. India, for example, is experiencing an alarming epidemic in T2D that threatens to sap the country's economic potency (S14).

Advances in medicine and technology offer some hope to those with type 1 diabetes — an autoimmune disorder that requires routine insulin injections. Immunomodulator agents under development could stop the body's misguided attack on the insulin-producing pancreatic cells (S4). And computer-controlled devices that monitor blood sugar levels and deliver insulin in response are taking some of the guesswork and inconvenience out of this vitally important task.

There is remarkably little certainty on how these conditions arise (S10). And although it remains unclear what triggers either T1D or T2D, the bacteria that live within us are implicated (S12).

Is diabetes preventable? On this question, the differences between T1D and T2D are perhaps most apparent (S18). Vaccines might one day be able to guard against T1D, but that day is still distant. T2D, on the other hand, appears to offer ample opportunity for individuals to manage their destiny through a healthy diet and exercise.

We acknowledge the financial support of Eli Lilly and Company in producing this *Outlook*. As always, *Nature* has full responsibility for all editorial content.

Herb Brody
Supplements Editor

CONTENTS

S2 STATISTICS
Diabetes in numbers
Disease burden and economic impact

S4 IMMUNOMODULATORS
Cell savers
Protect beta cells, treat the disease

S6 DEVICES
Managed by machine
The bionic pancreas goes automatic

S9 PERSPECTIVE
Rethink the immune connection
Carla Greenbaum

S10 PATHOLOGY
Cause and effect
Solving the puzzle of how diabetes arises

S12 MICROBIOME
The critters within
Evidence points to gut bacteria influencing the course of diabetes

S14 PUBLIC HEALTH
The Indian time bomb
Modernization has spawned a massive epidemic

S17 PERSPECTIVE
Testing failures
Thomas Mandrup-Poulsen

S18 PREVENTION
Nipped in the bud
Diet and exercise work — could a vaccine be in the offing?

COLLECTION

S21 **The NLRP3 inflammasome instigates obesity-induced inflammation and insulin resistance**
B. Vandanmagsar et al.

S31 **Solving the plot: early events are the key to diabetes intervention**
Alexander V. Chervonsky

S33 **Sleep and eating behavior in adults at risk for type 2 diabetes**
J. M. Kilkus et al.

S39 **The worldwide epidemiology of type 2 diabetes mellitus—present and future perspectives**
Lei Chen, Dianna J. Magliano and Paul Z. Zimmet

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the Nature Outlook Editorial guidelines available at http://www.nature.com/advertising/resources/pdf/outlook_guidelines.pdf

CITING THE OUTLOOK
Cite as a supplement to *Nature*, for example, *Nature* **Vol XXX**, No. XXXX Suppl, Sxx–Sxx (2012). To cite previously published articles from the collection, please use the original citation, which can be found at the start of each article.

VISIT THE OUTLOOK ONLINE
The *Nature Outlook Diabetes* supplement can be found at http://www.nature.com/nature/outlook/diabetes

All featured articles will be freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES
For UK/Europe (excluding Japan): Nature Publishing Group, Subscriptions, Brunel Road, Basingstoke, Hants, RG21 6XS, UK. Tel: +44 (0) 1256 329242. Subscriptions and customer services for Americas – including Canada, Latin America and the Caribbean: Nature Publishing Group, 75 Varick St, 9th floor, New York, NY 10013-1917, USA. Tel: +1 866 363 7860 (US/Canada) or +1 212 726 9223 (outside US/Canada). Japan/China/Korea:Nature Publishing Group — Asia-Pacific, Chiyoda Building 5-6th Floor, 2-37 Ichigaya Tamachi, Shinjuku-ku, Tokyo, 162-0843, Japan. Tel: +81 3 3267 8751.

CUSTOMER SERVICES
Feedback@nature.com

DIABETES IN NUMBERS

The number of people living with, and dying of, diabetes across the world is shocking: 90 million Chinese live with diabetes and 1.3 million died in 2011; 23% of Qatari adults have developed diabetes. Here we chart the extent of the global epidemic and present some of the implications for national governments by **Tony Scully**.

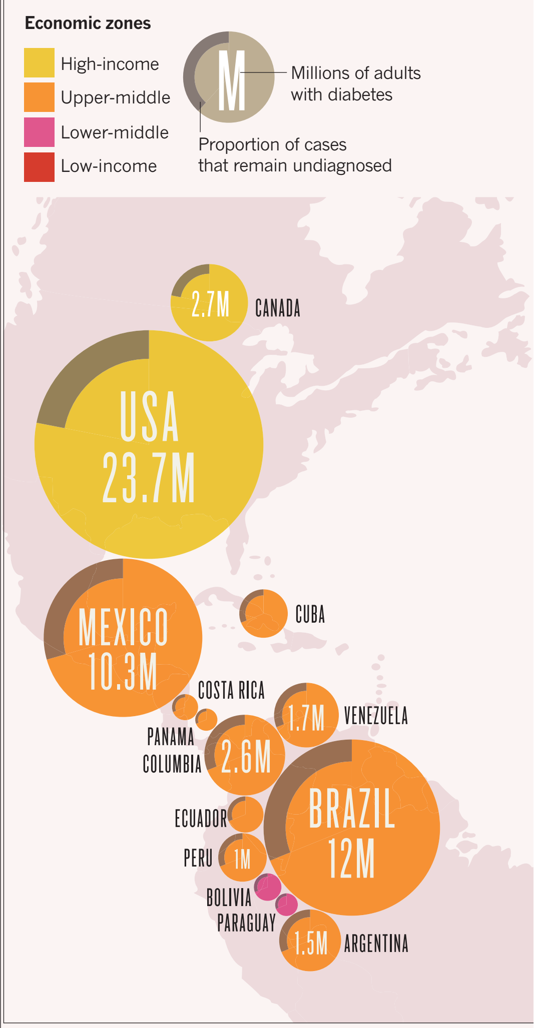
TSUNAMI OF DIABETES

1 Type 2 diabetes accounts for almost 90% of all cases of diabetes in adults worldwide. In general, as countries become richer, people eat a more sugar- and fat-rich diet and are less physical active — and the incidence of diabetes rises. On average, nearly 8% of adults living in high-income countries (see map for country classification) have diabetes. It is, however, upper-middle and middle-income countries that have the highest prevalence of diabetes; over 10% of adults in these countries have the condition.

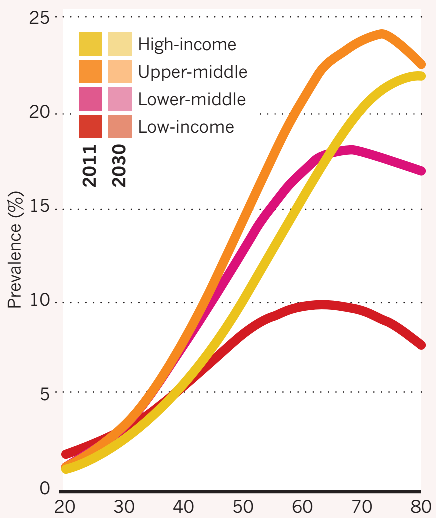
2 In high-income countries, diabetes primarily afflicts people over 50 years of age. But in middle-income countries, the highest prevalence is in younger people — the most productive age groups. As these people age, and as life expectancies increase, prevalence in older age groups will rise further. This trend will put a huge burden on healthcare systems and governments.

3 The mortality rate of diabetes varies sharply with the prosperity of the country. In 2011, the disease caused more than 3.5 million deaths in middle-income countries, of which more than 1 million were in China and just less than a million were in India. Approximately 1.2 adults die of a diabetes-associated illness per 1,000 cases in 2011 in low- and middle-income countries: more than double the mortality rate of high-income countries. Mortality rates are much lower in high-income countries with the greater healthcare recourses, but those tolls are still high: approximately 180,000 people died in the United States in 2011, for example.

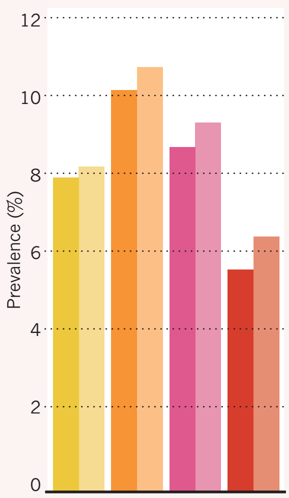
4 Unsurprisingly, high-income countries spent vastly more on diabetes-related costs in 2011 than lower-income countries. In developing countries, the looming costs in human lives, healthcare expenditure and lost productivity threatens to undo recent economic gains.



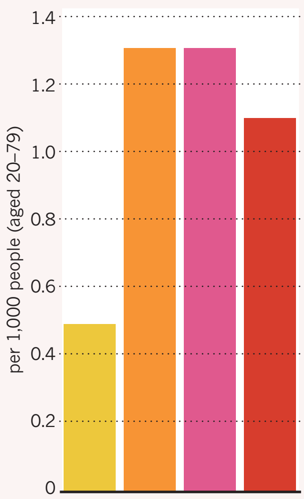
346 M people worldwide have diabetes. More than 80% of diabetes deaths occur in low- and middle-income countries, according to the WHO.



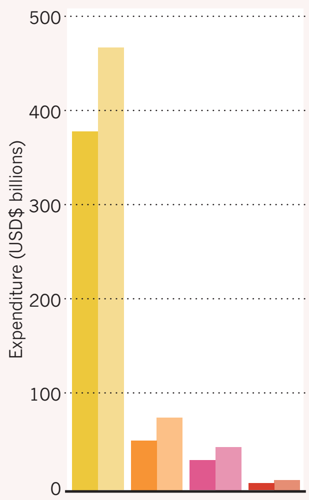
1 Prevalence of diabetes among adults (aged 20–79)



2 Prevalence of diabetes in adults (aged 20–79)



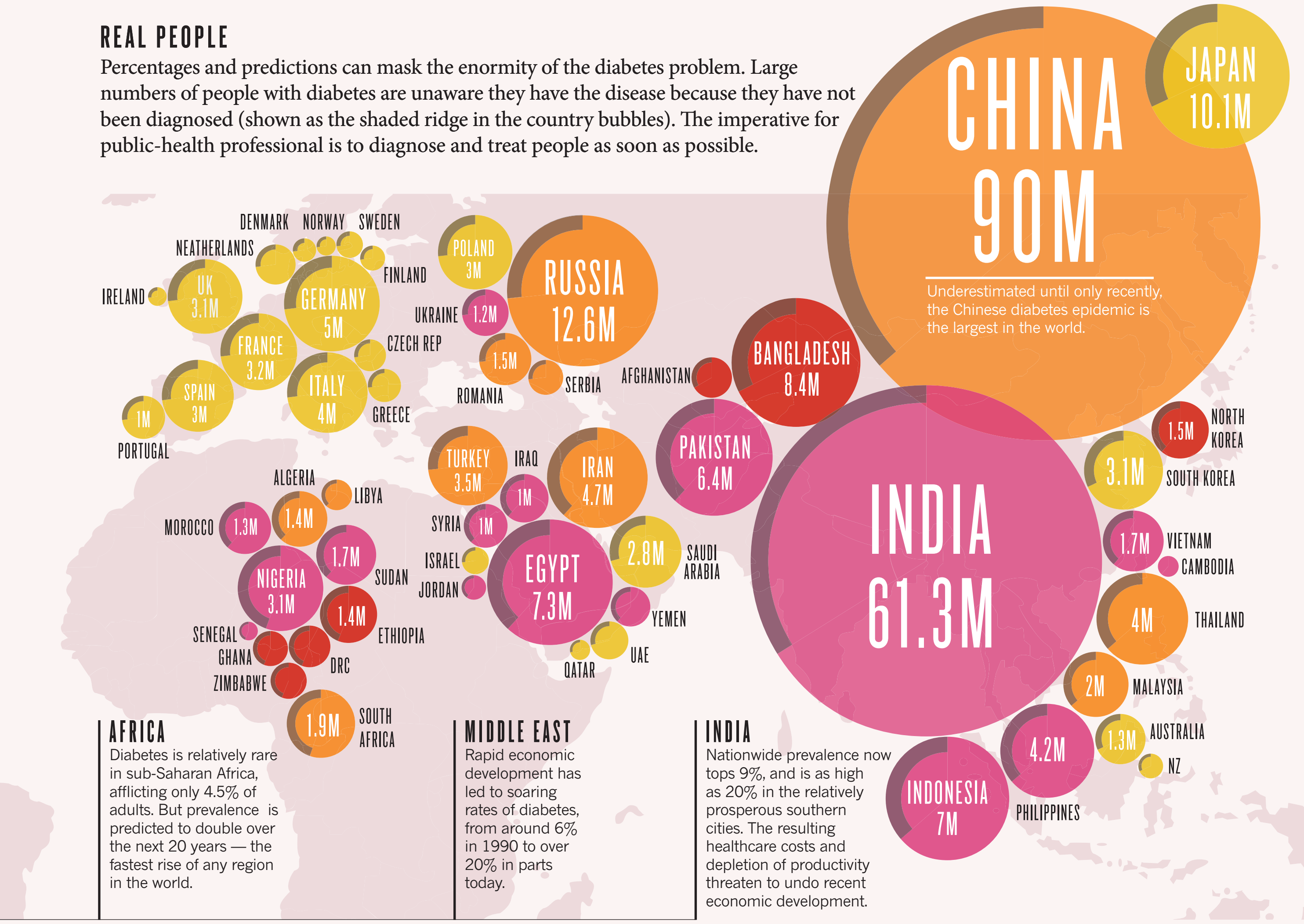
3 Mortality rates in adults (aged 20–79)



4 Total healthcare expenditures (2011)

REAL PEOPLE

Percentages and predictions can mask the enormity of the diabetes problem. Large numbers of people with diabetes are unaware they have the disease because they have not been diagnosed (shown as the shaded ridge in the country bubbles). The imperative for public-health professional is to diagnose and treat people as soon as possible.



AFRICA

Diabetes is relatively rare in sub-Saharan Africa, afflicting only 4.5% of adults. But prevalence is predicted to double over the next 20 years — the fastest rise of any region in the world.

MIDDLE EAST

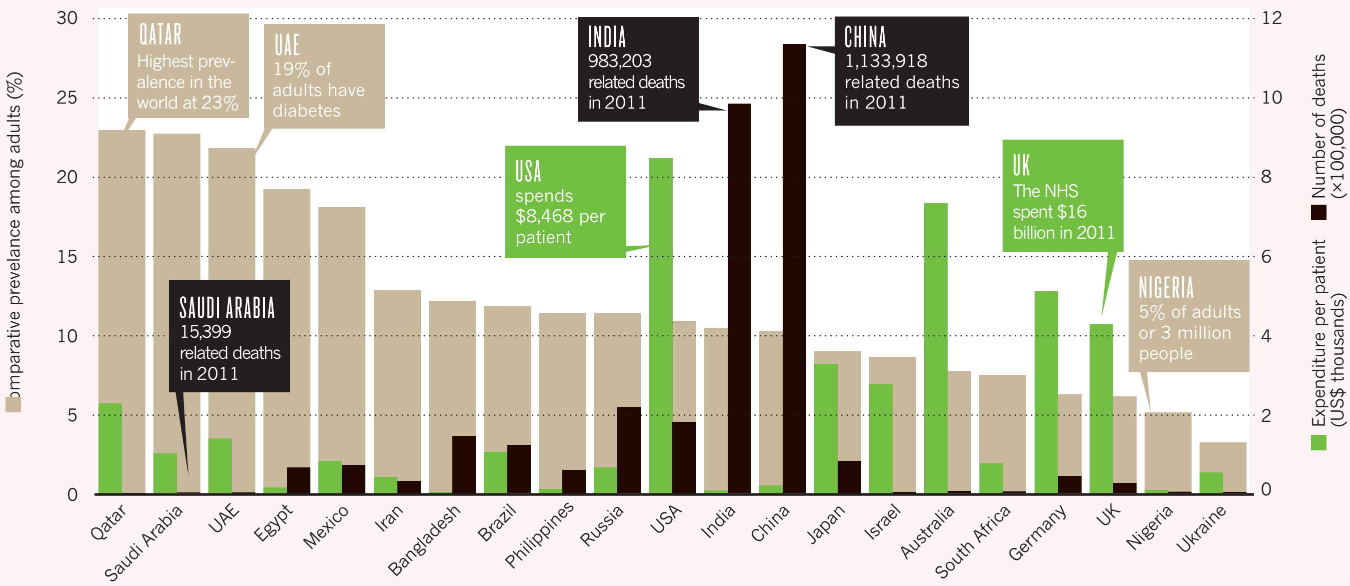
Rapid economic development has led to soaring rates of diabetes, from around 6% in 1990 to over 20% in parts today.

INDIA

Nationwide prevalence now tops 9%, and is as high as 20% in the relatively prosperous southern cities. The resulting healthcare costs and depletion of productivity threaten to undo recent economic development.

THE INVESTMENT GULF

Figures for a selection of countries detail national prevalence alongside total expenditure per patient and number of diabetes-related deaths. The countries with the highest prevalence and rates of mortality spend far less per patient than some other countries. As epidemics mature, costs and mortality are estimated to rise.



International Diabetes Federation. The IDF Diabetes Atlas. Fifth Edition. Brussels (2011)..

PHOTO PROVIDED BY VANDERBILT UNIVERSITY MEDICAL CENTER

Kerby Bennett is a participant in the NIH's TrialNet diabetes prevention study.

IMMUNOMODULATORS

Cell savers

In type 1 diabetes, the immune system goes haywire and depletes insulin-producing cells. Drugs that interfere with this process could one day reverse the disease's course.

BY SARAH DEWEERDT

Heart attack, stroke, aneurysm: each is a potentially fatal event. Most people probably wouldn't consider type 1 diabetes to be as urgent, but at least one researcher argues that it is.

Jean-Francois Bach, an immunologist at the University of Paris-Descartes, argues that type 1 diabetes (T1D) should be considered a "medical emergency", and that the goal of treatment should be to reverse the disease as soon as possible after diagnosis.

T1D occurs when the immune system mistakenly attacks beta cells — the insulin-producing cells in the pancreas. The assault takes years before the disease manifests, but only one-third of beta cells survive in most patients by the time diabetes is diagnosed, so it is critical to save those beta cells that remain. If the immune attack could be halted and those remaining cells could be preserved, scientists believe that they could be sufficient to produce most of the body's insulin on their own.

Over the past decade, numerous clinical trials have tried to use immune-modifying drugs, many borrowed from the treatment of other autoimmune diseases, to try to save beta cells. So far, none have worked especially well. And because the disease is most often diagnosed in children and young adults, any effective treatment would have to be tolerated over the course of many years. But researchers are continuing to hone the dosage, timing and perhaps combination of drugs to a therapy that works.

HIGH HOPES

In the 1980s, studies demonstrated that suppressing the immune system of people recently diagnosed with T1D reduced their insulin dependence and provided persuasive evidence that T1D is an autoimmune disease. "Back then it wasn't so clear," says immunologist Jeffrey Bluestone of the University of California, San Francisco.

The early immunomodulators, such as cyclosporine and antithymocyte globulin, were blunt instruments, targeting not just the immune cells responsible for killing the beta cells, but other parts of the immune system as well. The drugs were too toxic for patients to take for extended periods, and any protective effect failed to persist after treatment. But even this limited effectiveness was enough to spur researchers to look for more specific immune modulators that would affect the precise mechanisms of beta-cell destruction in diabetes.

Much attention has focused on T cells, thought to be behind the targeted killing of beta cells in T1D. Antibodies against CD3, a receptor found on T cells, can prevent or even permanently reverse diabetes in non-obese diabetic (NOD) mice and other mouse models of T1D. "What was exciting in the mouse model was that we gave short-term dosing and we got long-term effects," says Bluestone. He and other researchers hope that anti-CD3 might attenuate the immune system rather than switch it off with anti-rejection drugs.

Two different anti-CD3 antibody drugs, teplizumab and otelixizumab, have been tested in humans. Studies in Europe and the United States showed that a short course of treatment — two weeks or less — in people recently diagnosed with diabetes can improve beta-cell function for as long as five years.

In contrast to the mouse results, the effect in humans is temporary; eventually, beta-cell depletion begins again and the disease progresses. A bigger setback came in 2011, with the results of two large, phase III clinical trials of these agents on recently diagnosed patients. Although the design and endpoints of the studies were slightly different, both showed that the anti-CD3 approach did not improve on standard insulin therapy after one year[1].

That's not an unfamiliar result. Indeed, Bluestone says, showing improvement over standard therapy is "really tough in the first year after diagnosis". That's because insulin formulations and delivery systems are now so good that most patients can easily keep their diabetes well controlled at first. Comparing any immunosuppression strategies to insulin treatment has been the undoing of a number of other drug treatments, he says.

The same story of high hopes dashed in phase III trials unfolded for glutamic acid decarboxylase (GAD), a molecule normally found in the pancreas and one of the antigens targeted for destruction in the immune attack. Delivering GAD in a different form might help the immune system regard the molecule as friend rather than foe. But a European study found that injections of GAD failed to either stem the loss of beta-cell function or improve diabetes control over the course of 15 months in recently diagnosed T1D patients[2]. A similar trial in the United States was halted because GAD didn't seem to be effective.

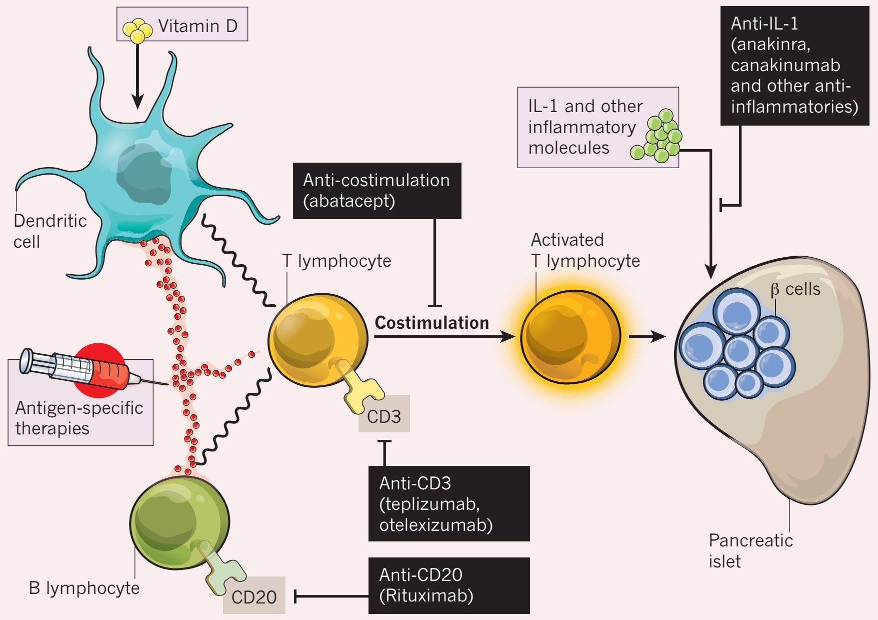
Jay Skyler, chair of Type 1 Diabetes TrialNet, a National Institutes of Health-funded research consortium in the United States, says that the problem with GAD might have been its delivery. Animal studies showed some promise when the molecule was administered orally, nasally or by

injection into the abdominal cavity, but both the US and European clinical trials used an injection under the skin. "So I personally am not ready to give up on GAD," says Skyler, who is also a professor of medicine and paediatrics at the University of Miami in Florida. "I want to see more data."

There is some optimism about other antigens, such as insulin and its precursor, proinsulin, which might form the basis of other immune-modulating therapies. This approach offers the advantage of specificity: rather than altering the function of whole classes of immune cells, which might render a patient susceptible to infection, antigen therapies affect only activated T cells that attack beta cells. This precision is "a very important safety net", says Chantal Mathieu, an endocrinologist at the Catholic University of Leuven in Belgium.

IMMUNOTHERAPY

The object in each case is to prevent the immune system (T cells) from attacking the beta cells in the pancreatic islet (shown on the right).



DIVERSE TWEAKS

Mathieu is investigating other, gentler strategies for immune modulation. She is coordinator of Natural Immunomodulators as Novel Immunotherapies for Type 1 Diabetes (NAIMIT), an EU-funded project. For example, Mathieu's group has previously shown that vitamin D can prevent diabetes in NOD mice. "The problem is that the doses we needed were huge," she says — equivalent to 1,000 times what could be used in humans. Such massive amounts of vitamin D can cause heart arrhythmias, kidney stones and dangerously high levels of calcium.

But researchers have not written off vitamin D just yet. A team led by diabetes and immunology specialist Bart Roep at Leiden University in the Netherlands showed that when dendritic cells (a type of immune cell) mature *in vitro* in the presence of vitamin D, they dampen T-cell responses. "This is a very elegant way of exploiting the immune-modulatory effect of vitamin D," Mathieu says. The team hopes to begin human studies, infusing the dendritic cells back into T1D patients, sometime in 2012.

Another possible immunological tweak is to block costimulation — the final stage of T-cell activation. A TrialNet study taking this approach enrolled 77 people with recently diagnosed diabetes. They each received 27 infusions of abatacept, a costimulation inhibitor used in rheumatoid arthritis, over 2 years. Results were mixed. Those who received abatacept had better beta-cell function than the control group[3]. However, the drug had its strongest effects during the first 6 months of treatment. After that, the abatacept group lost beta-cell mass and function at the same rate as controls.

> ***"Immunotherapy could extend beyond treatment to become the basis for diabetes prevention."***

TrialNet researchers found similar results in a study targeting B cells, which help T cells recognize antigens. They administered four doses of rituximab, an antibody against the CD20 receptor found on B cells, to 57 people with recently diagnosed diabetes. A year later, those who had received rituximab had better beta-cell function and required less insulin than controls[4]. But again, the effects were most dramatic early on, and by 6 months the rituximab group were also losing beta cells. "These drugs all seem to have similar effects, and they all seem to have effects only during this certain window of time — within that first 3–6 months," says Carla Greenbaum, director of the diabetes programme at Benaroya Research Institute in Seattle, Washington, and vice-chair of TrialNet.

INNATE INTERPLAY

This research cul-de-sac has prompted researchers to explore more than just the antigen-specific responses of B and T cells (that is, adaptive immunity) — and to look at the non-specific, or innate, immune responses such as inflammation, as well. As early as the mid-1980s, endocrinologist Mandrup-Poulsen, now at the University of Copenhagen in Denmark, showed that one molecule involved in innate immunity — the pro-inflammatory interleukin-1 (IL-1) — can kill insulin-producing beta cells. The recent revival of interest in innate immunity has prompted further study. In 2011, a preliminary study found that 15 children recently diagnosed with T1D who were given a 28-day course of the IL-1 blocker anakinra needed less insulin 4 months after diagnosis than did controls[5].

Two phase II trials of IL-1 blockers in T1D are just wrapping up. A European study headed by Mandrup-Poulsen also uses anakinra; a TrialNet study in the United States uses a similar drug called canakinumab. Both groups plan to announce the results of their studies at the American Diabetes Association meeting in Philadelphia, Pennsylvania, June 2012. Even if these studies are negative, Mandrup-Poulsen argues, researchers should consider investigating IL-1 blockers as part of combination therapy.

In fact, a recent mouse study provides support for such a strategy. Researchers found that combining anti-CD3 antibody with anakinra — at doses too low for either drug to work alone — permanently reversed diabetes in NOD mice[6]. "This is really exciting because it would indicate that if you titrate these two immune modulatory agents you may obtain very potent effects on the disease process," says Mandrup-Poulsen, a co-author of the study's report.

The impact of immunotherapy could extend beyond treatment. It might also become the basis for diabetes prevention, because the immune-related destruction of beta cells is thought to begin several years before diabetes becomes clinically apparent. "In theory, stopping the immune attack before it's fully blown ought to be more effective than stopping it when everything is underway," says Skyler. In other words, the best strategy for dealing with an immune emergency would be to prevent it altogether. ■

Sarah DeWeerdt *is a science writer based in Seattle, Washington.*

1. Sherry, N. *et al. Lancet* **378,** 487–497(2011).
2. Ludvigsson, J. *et al. N. Engl. J. Med.* **366,** 433–442 (2012).
3. Orban, T. *et al. Lancet* **378,** 412–419 (2011).
4. Pescovitz, M. *et al. N. Engl. J. Med.* **361,** 2143–2152 (2009).
5. Sumpter, K. M. *et al. Pediatric Diabetes* **12,** 656–667 (2011).
6. Ablamunits, V. *et al. Diabetes* **61,** 145–154 (2012).

Rethink the immune connection

Recent research suggests that the fight against type 1 diabetes is focusing too narrowly on the adaptive immune system, says **Carla Greenbaum.**

Decades ago, investigators established the pathology of type 1 diabetes (T1D) — the adaptive immune system mistakenly attacks the insulin-producing beta cells in the pancreatic islets. Long before the clinical onset of disease (defined by hyperglycaemia), the immune assault triggers a progressive process of beta-cell dysfunction and cell death. As this process unfolds, diabetes-related autoantibodies begin to circulate through the body, and the secretion of insulin is impaired. The attack continues after diagnosis.

This model has served us well in predicting who will get the disease. For example, a relative of someone with T1D who has one of the diabetes-related antibodies has about a 3% chance of developing T1D over the next five years; those with two or more antibodies have a 35–85% chance. Although some beta cells remain when clinical symptoms appear, over time the beta cells are completely destroyed.

But an explosion of data about the immune system is yet to yield a cure or prevention strategy for T1D. And we now have the results of several clinical trials testing the hypothesis that it is the adaptive immune system that is wreaking havoc on beta cells. In individuals with recently diagnosed diabetes, altering components of the adaptive immune system, for example through anti-T-cell therapies or anti-B-cell therapies, seems to improve insulin secretion (an indication of beta-cell function) by roughly 25% compared to control subjects[1]. Better beta-cell function is associated with important clinical benefits — less hypoglycaemia and fewer complications — but with limited clinical data, the long-term benefits to individual patients remain unknown.

BEFORE WE EMBARK ON OTHER LARGE CLINICAL TRIALS, WE NEED MORE BASIC RESEARCH, PARTICULARLY PROOF-OF-MECHANISM STUDIES.

UNSUSTAINABLE RESPONSE

Moreover, attempts to use short-term treatments to induce long-term immune tolerance of beta cells in a bid to stop disease have not worked. The best interventions so far have slowed the rate of decline of beta-cell function within the first months of diagnosis, but repeated or continued treatments failed to sustain this response. It has also been postulated that treating a T1D patient with antigen, such as insulin or GAD65 in alum, could safely induce tolerance (see ‘Cell savers’, page S4). However, beta-cell function continues to deteriorate in people with diagnosed diabetes who receive antigen. It is possible that antigen therapy might work at different doses, in different populations of people (particularly earlier in the disease course), and in conjunction with other therapies. And yet, clinical trials testing insulin as a prophylactic — whether delivered nasally or parenterally — also failed to prevent diabetes in those who were identified as at risk for type 1 diabetes[2,3].

One interpretation of these clinical failures is that we have not been aggressive enough in our attempts to save beta cells in those with T1D. An uncontrolled trial of haematopoietic stem-cell therapy to save beta cells has had some success[4]. However, this approach has unknown benefits and is fraught with risks, such as pneumonia and decreased gonad function; just because we can do it, it doesn’t mean we should — especially in a disease affecting children.

Rather than not being aggressive enough with therapies, an alternative explanation to the limited success seen to date is that we have narrowly defined therapeutic targets in our intervention trials — namely molecules and pathways of the adaptive immune system. There is undoubtedly a role of the innate immune system and inflammation in beta-cell destruction; clinical trials testing this hypothesis, including blocking the proinflammatory protein interleukin-1, are underway. Moreover, beta cells in T1D might not be the victim of an immune attack, but rather have defective responses to injury or stress. It is true that genome wide association studies (GWAS) have implicated immune related genes. But these same genes have other functions, including influencing beta-cell function and response. When we look at results from an immune-centric approach, we risk missing other factors that can contribute to beta-cell dysfunction. For instance, several hypotheses suggesting that environmental and behaviours factors play a role in the climbing incidence of T1D world wide[5] await further testing. Our next generation of trials must address multiple components — immunology, genetics, environment and behaviour. Animal models alone will not be enough to guide our future endeavours.

UNSUSTAINABLE RESPONSE

Before we embark on other large clinical trials, we need more basic research, particularly proof-of-mechanism studies[6]. Such clinical research entails testing new therapeutic approaches in a small number of individuals to measure a biological or mechanistic response. This is the way to examine how alterations in metabolic state or beta-cell stress affect immune function, or to assess off-target effects of combination therapy. Evaluating all data should allow us to guard against evidentiary conservatism (the tendency to base clinical inferences on narrow classes of evidence) and to design the next generation of studies with open minds. To change the course of diabetes, we might need to alter our course. ■

Carla Greenbaum *is an endocrinologist and director of the diabetes program at the Benaroya Research Institute in Seattle, Washington.*
email: cjgreen@benaroyaresearch.org

1. Orban,T. *et al. Lancet* **378,** 412–419 (2011).
2. Diabetes Prevention Trial Study Group. *N. Engl. J. Med.* **346,** 1685–1691 (2002).
3. Nanto-Salonen, K. *et al. Lancet* **372,** 1746–1755 (2008).
4. Voltarelli, J. C. *et al. J. Am. Med. Assoc.* **297,** 1568–1576 (2007).
5. Karvonen, M. *et al. Diabetes Care* **23,**1516–1526 (2000)
6. Kimmelman, J. & London, A. J. *PLoS Med.* **8,** e1001010 (2011).

COURTESY OF MEDTRONIC, INC.

Software may soon close the loop between glucose sensors (left) and insulin pumps (right).

MEDICAL DEVICES

Managed by machine

Artificial pancreases promise to take the decision-making — and human mistakes — out of managing type 1 diabetes.

BY ELIE DOLGIN

Leah Moynihan lifts her shirt to reveal half a dozen devices strapped to her midriff: four glucose sensors and two hormone pumps attached to her belly and a pack of remote controls slung across her chest. On this Saturday in February, Moynihan is wired up to test what could be a major advance in the treatment of type 1 diabetes (T1D): a bionic pancreas that automatically dispenses the right amount of insulin in response to fluctuations in blood glucose levels.

"It looks like chewing gum and paper clips right now," admits Kendra Magyar, a research nurse at Massachusetts General Hospital (MGH) in Boston who has type 1 diabetes herself and is helping to run the trial. "If it all works out, it'll get smaller and less invasive."

The current standard of care for treating T1D leaves much to be desired. Typically, people prick their fingers several times a day to monitor their blood sugar levels. They then try to regulate their blood glucose, either by eating sugary foods if blood sugar levels are low (hypoglycaemia) or by injecting themselves with insulin when glucose levels spike. Two types of wearable devices that help people manage the condition have recently hit the market. One, the continuous glucose monitor, is a tiny sensor placed just under the skin that checks sugar levels automatically every few minutes. The other is an insulin pump about the size of a mobile phone that attaches to a fine needle implanted under the skin to deliver the missing pancreatic hormone at the click of the button.

The trouble is that both systems still require people to decide for themselves if, when and how to get their blood sugar levels back into the normal range — and a wrong decision can be deadly. People with T1D suffer an average of two episodes of symptomatic hypoglycaemia per week, and as many as 10% of deaths in this patient group are caused by insulin-related complications. "When you look at the care and the burden of type 1 diabetes — testing and correcting 24 hours a day — it really is unbelievable," says Dana Ball, programme director for the T1D programme at the Helmsley Charitable Trust, a New York-based non-profit organization that is partly funding the MGH trial. "A more sophisticated device would be incredible."

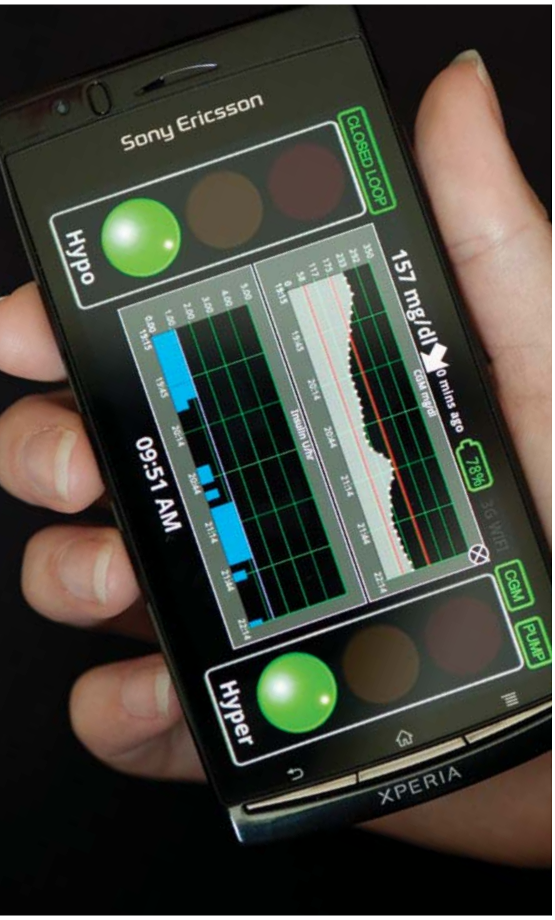
For Moynihan, a nurse practitioner at the nearby Mount Auburn Hospital in Cambridge, Massachusetts, who has lived with T1D for close to three decades, such a device could not come soon enough. Diabetes "interferes with my life every day, all day long", she says. "So it gives me some hope that there will be more than my having to think about how much insulin I need to take or whether I need a snack."

IN THE LOOP

To improve the quality of life for people with T1D and help prevent diabetes-related premature death, several researchers are designing automated systems that close the loop between glucose monitors and insulin infusion devices by transmitting information wirelessly from a sensor to an insulin pump. These closed-loop artificial pancreases rely on advanced control algorithms — mathematical formulations run on software — to make therapeutic decisions and accurately regulate blood sugar in real time with minimal human input. "This is an unprecedented kind of technology in which you're handing over therapeutic decisions to software," says Ed Damiano, a biomedical engineer at Boston University in Massachusetts involved with the MGH trial. "As soon as it's available, it will make the current standard of care obsolete."

Artificial pancreases trace their roots back more than 35 years to the Biostator, a device introduced by Indiana-based Miles Laboratories in the late 1970s. The refrigerator-sized controller relied on intravenous blood readings and intravenous infusions of insulin, which meant its use was limited to the hospital. Still, the Biostator proved that such a closed-loop platform was possible. Products that were more portable soon followed, but the surgical procedures needed to implant the sensors and

➲ **NATURE.COM**
Nutrition & Diabetes, a new open-access journal:
go.nature.com/oypqa7



Closed-loop algorithms can run on smartphones; this version shows insulin delivery and glucose level.

MELODY KOMYEROV FOR BOSTON UNIVERSITY

pumps deep inside the body, among other safety and usability problems, prevented their broad commercialization.

Closed-loop devices for use in the home took a big step forwards about ten years ago, with the roll-out of glucose sensors that could be implanted beneath the skin. The first-generation devices provided only retrospective data that could be analysed after the fact to inform disease management. But in 2005, Medtronic, a medical device company based in Minneapolis, Minnesota, began selling the Guardian RT, which relayed glucose results every five minutes. Subcutaneous devices had already been available for insulin delivery for decades. Now all that was missing was the link between the two.

IT TAKES TWO

To spur the development of algorithms that could make therapeutic decisions, the JDRF, a New York-based non-profit foundation, initiated the Artificial Pancreas Project in 2005. This multimillion dollar initiative brought together diabetes researchers and businesses determined to make the artificial pancreas a reality. Around the same time, the US Food and Drug Administration (FDA) identified the artificial pancreas as a top priority and, together with the US National Institutes of Health, formed the Interagency Artificial Pancreas Working Group to identify and work through any clinical and scientific challenges. Meanwhile, government funding bodies in the United States and Europe, as well as many medical device companies, started spending tens of millions of dollars to encourage the development of an artificial pancreas.

In the wake of rapid progress, a handful of independent research groups launched human clinical trials, and several algorithms are being tested (see 'Control issue'). For the most part, studies have been conducted under the controlled confines of the hospital setting, often with participants hooked up to laptop computers and intravenous backup systems that limit their mobility, as Moynihan was. But some investigators have taken their devices to the next level.

At the Princess Margaret Hospital for Children in Perth, Australia, Medtronic is running its algorithm on a BlackBerry smartphone. In Italy and France, researchers are using mobile phones and tablet computers to conduct trials in hotels — not hospitals — with doctors and engineers in separate rooms in case safety problems arise. "The patients wanted to go home with it," says Eric Renard, a diabetes specialist at Montpellier University Hospital in France who is leading the hotel-based trial. "After only a few hours, they say they're completely different. Never before have they had this feeling that they don't have to think about their disease." In March 2012, the FDA approved a similar trial using the same technology at the University of Virginia in Charlottesville and at the Sansum Diabetes Research Institute (SDRI) in Santa Barbara, California.

In the United States, some investigators have also started experimenting with systems that try to improve how the artificial pancreas works. For example, Damiano's team and an independent group in Portland, Oregon, are using a pancreatic hormone called glucagon to help raise blood glucose when too much insulin has been delivered and blood sugar levels start to plummet. At Yale University School of Medicine in New Haven, Connecticut, researchers are adding pramlintide, a synthetic version of another human hormone called amylin, to help slow the absorption of nutrients from the gut as glucose levels rise after mealtimes. The Yale group has also tested a patch that heats the skin before insulin release to increase blood flow to the site in order to speed up the hormone's uptake. Given the inherent lag times associated with subcutaneous insulin absorption, "you're going to have a problem with catch up", says Stuart Weinzimer, a paediatric endocrinologist who is leading the Yale trials. "Anything you can do to speed up insulin delivery or slow down glucose absorption will help."

> ***"Never before have patients had this feeling that they don't have to think about their disease."***

UNIVERSITY OF VIRGINIA PATENT FOUNDATION

A RISKY PROPOSITION

Although developers of artificial pancreases have differing opinions about the best closed-loop design, all agree that safety must remain a top priority as more authority is handed over to the device. "Hypoglycaemia is extraordinarily dangerous. You lose consciousness and then you have seizures and you die if someone doesn't help you," warns Steven Russell, a diabetes specialist at MGH who is collaborating with Damiano on the trials in Boston. "Giving over control entirely to a machine is a high-risk proposition," he says, making it imperative that the process be "done properly".

To help make the safe transition to a fully closed-loop system that requires minimal human input, many experts and companies are advancing hybrid control algorithms that are only partly automated. "We want to take iterative steps to closing the loop," says John Mastrototaro, vice-president of global medical, scientific and health affairs at Medtronic's diabetes division in Northridge, California.

The first such product could be Medtronic's Paradigm Veo, an insulin pump that automatically turns off when a sensor reports that glucose levels have fallen below a certain level. Already available in Europe, this 'low glucose suspend' system is now undergoing in-home testing in the United States, and is expected to receive regulatory approval in 2013.

Subsequent partly automated systems will probably benefit from technological improvements. The next logical step is a predictive low-glucose sensor that anticipates declining

Ed Damiano checks the readings on his son David's continuous glucose monitor.

glucose levels, rather than relying on a hard cut-off point as the Paradigm Veo does. Then maybe there will be a device that automatically increases the insulin rate when blood glucose levels rise above a certain threshold, followed perhaps by a fully closed-loop system that only works when people are asleep, thereby avoiding

CONTROL ISSUE
The algorithm method

Diabetes researches and clinicians generally agree that a safe and effective artificial pancreas should provide better treatment than the current standard of care. But the bioengineers behind the systems don't agree on the best type of algorithm to control the closed-loop devices.

In one camp sit the advocates of so-called 'proportional-integral-derivative' (PID) controllers, a simple strategy widely used in feedback control in settings ranging from missile steering to automobile cruise control. For artificial pancreases, PID-based algorithms use glucose values and rates of change to make calculations of insulin dosing. Gary Steil, a former Medtronic engineer who is now developing his own algorithms and running clinical trials at Children's Hospital Boston in Massachusetts, says that the PID approach best emulates how the body's insulin-producing beta cells manage glucose naturally, as they simply react to blood glucose levels and then spit out hormones as needed. "Everything in this algorithm is linked to something that the beta cell does," Steil says.

But others endorse a more predictive strategy to make up for the unavoidable time lags associated with subcutaneous glucose sensing and insulin release. Known as 'model-predictive control', this method tries to plan several moves ahead in someone's glucose control based on past actions and responses. According to Boris Kovatchev, director of the University of Virginia Center for Diabetes Technology in Charlottesville, this level of built-in prediction is vital to ensure patient safety. "Safety cannot be reactive," he says. "It's too late to be reactive."

Some researchers, however, say this whole debate around control theory techniques is a red herring. "Algorithms aren't the issue," says Ken Ward, an endocrinologist at Oregon Health and Science University in Portland. "If we had a really reliable sensor and reliably fast insulin, I think the artificial pancreas would work with any number of algorithms." — *E. D.*

ELIE DOLGIN

Leah Moynihan gives her bionic pancreas a test ride.

the confounding factors of meals, stress and exercise, all of which can complicate blood glucose management. Importantly, all these hybrid devices would be automated some of the time but still maintain some degree of human input in the intervening periods.

"This is the Wright Brothers at Kitty Hawk when everyone wants to go to the Moon," says endocrinologist David Klonoff, medical director of the Diabetes Research Institute at Mills-Peninsula Health Services in San Mateo, California, who has been involved in the Paradigm Veo's in-hospital trials. "It's a first step, but you've got to start somewhere."

"We all have this goal of a fully automated system," says Howard Zisser, director of clinical research and diabetes technology at the SDRI. "But we need to harvest some of this low-hanging fruit," he says of semi-automated systems, which can be put into practice more easily. What's more, he adds, "it will be easier to convince the regulatory authorities that an artificial pancreas can readily help people with type 1 diabetes."

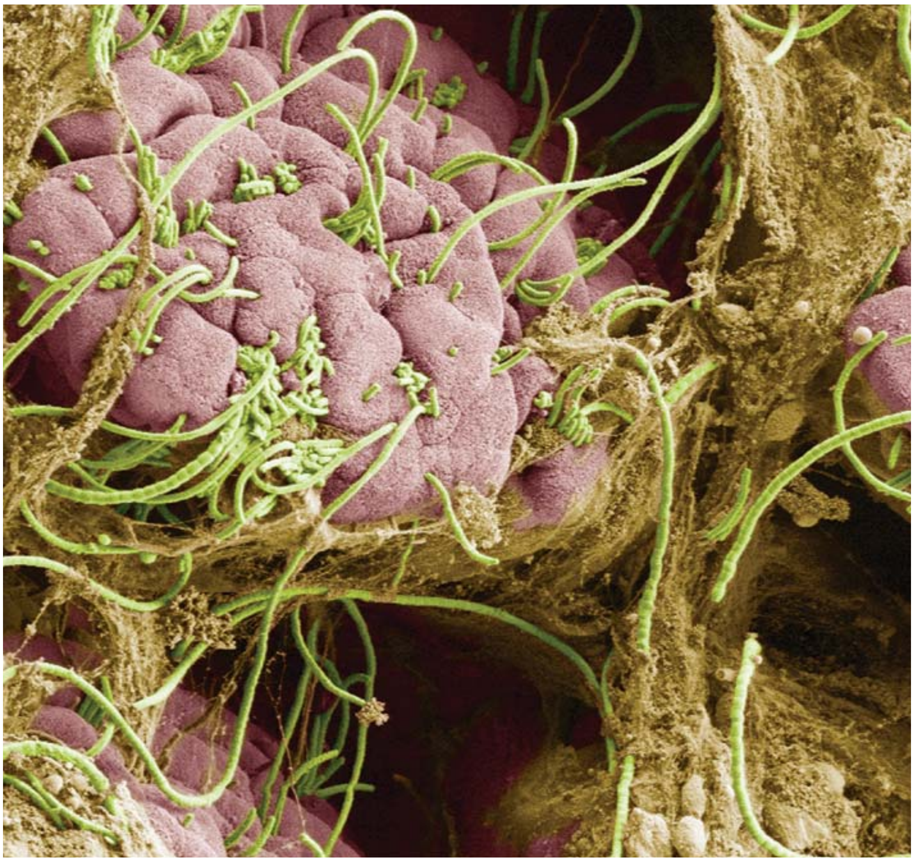
Getting to that point, however, could be a long and bumpy road, especially in the United States — as shown by the slow path to approval for low glucose suspend systems. To speed the approval process along, in October 2011 the JDRF launched a campaign to convince the FDA, which was drawing up guidelines on artificial pancreases at the time, to create a clear and reasonable path to approval for closed-loop devices. "The reason we're putting pressure on here is because there is a critical unmet medical need," says Aaron Kowalski, research director of the JDRF's Artificial Pancreas Project. "We all want safe and effective products, but we also appreciate that people with diabetes are struggling now and the technology exists to help them do better."

The response to the JDRF appeal was overwhelming. In only three weeks, more than 100,000 people signed a petition — and the FDA paid attention. In December 2011, the agency released draft guidelines in which it promised to be flexible on trial sizes, durations and clinical endpoints needed for approval of an artificial pancreas. "There is no magic number or glucose level that the FDA believes is necessary to approve these devices," says Charles Zimliki, chair of the FDA's Artificial Pancreas Critical Path Initiative. "We're really trying to say: 'Come in and talk to us as you're developing these systems.'"

In February 2012, Damiano and Russell did just that. They met with FDA officials to discuss setting up five-day trials at MGH with their algorithm running on an iPhone. "Subjects will have free run of the entire hospital campus," Damiano says. After that, they hope to run 12-day trials involving MGH staff with type 1 diabetes; these study volunteers would go about their jobs at the hospital as normal while wired up to the device, and would even be able to sleep at home while still connected. Then, the Boston team plans to conduct one- to two-week trials with children at diabetes camps, followed sometime in 2014 by pivotal long-term outpatient trials.

Through it all, Damiano remains confident that a fully closed-loop device will make it to market sometime before his son, who was diagnosed with T1D in 2000 at just 11 months of age, graduates from high school. "Before my son goes to college, he has to wear one of these things," he says. "Or else I'm going with him." ■

Elie Dolgin *is a news editor with* Nature Medicine *in Cambridge, Massachusetts.*



Segmented filamentous bacteria (SFB) in the terminal ileum of an 8-week old Taconic B6 mouse.

ALICE LIANG, DOUG WEI, ERIC ROTH, NYU.

MICROBIOME

The critters within

Your gut microflora might be aiding and abetting diabetes.

BY LAUREN GRAVITZ

In 2004, Fredrik Bäckhed and his colleagues at Washington University in St Louis, Missouri, noticed that gnotobiotic mice — born and raised to be free of germs — tended to be slimmer than their conventional counterparts. After they transplanted the feces of normal mice to germ-free ones, the rodents gained weight and their insulin was less effective at lowering blood sugar levels[1]. Some of the same researchers later transplanted bacteria from the intestines of either lean or obese mice into the guts of gnotobiotic mice; those animals that received bacteria from obese mice gained nearly twice as much weight as mice on the same diet that received bacteria from lean donors[2]. These studies jump-started research that is transforming the way we think about obesity and diabetes.

The average human gut is home to trillions of bacteria. They outnumber the cells of their human host by a factor of ten to one, and collectively their genes outnumber human genes one hundred-fold. Together, they function as another organ, complementing and interacting with human metabolism in ways not fully understood. But one thing is becoming clear: the composition of bacterial species in the gut can influence the course of diabetes and its treatment.

"I have been studying diabetes for the past 25 years, and this is the most important discovery that has been made in my field," says Rémy Burcelin, research director at the French National Medical Research Institute (INSERM) in Toulouse. "We've discovered a new organ. We know there is a brain, a pancreas, a liver. Now we also know there are microbiota."

Humans and the microbiome — the bacteria that reside in and on us — have co-evolved for millennia. But lately we have been messing with the delicate balance between our flora and ourselves by eating more fats and sugars, by washing with antibacterial soap, and by taking antibiotics at the faintest hint of infection. This shift in behaviour has coincided with an increase in the incidence of type 1 and type 2 diabetes, both of which are rising at a pace that cannot be down to genetics alone (see 'Cause and effect' page S10).

"There's an order of magnitude more bugs in our gut then there are cells in our bodies, so it's not very difficult to imagine that they would have a profound impact on metabolic balance and metabolic activity," says Christopher Newgard, a metabolism researcher at Duke University in Durham, North Carolina. "But, as attractive and enticing as the theory may be, it has not yet been proven in a systematic way."

FINDING A FOOTHOLD

Researchers know that certain phyla of bacteria are more populous in obese mice, whereas others are more common in lean ones, and the same seems to hold true in people. Moreover, bacterial composition in the gut can improve or worsen insulin resistance in mice and, initial results suggest, in people. There also appears to be a connection between inflammation and the development of insulin resistance — some of the bacteria in obese and insulin-resistant people have the potential to trigger chronic, low-grade inflammation. What researchers don't know is how all these pieces fit together.

Two questions loom large. First, what is cause and what is effect? That is, do altered bacterial populations trigger insulin resistance or are they the product of something else in the body — and to what extent does an atypical microbiome affect the metabolism of it human host?

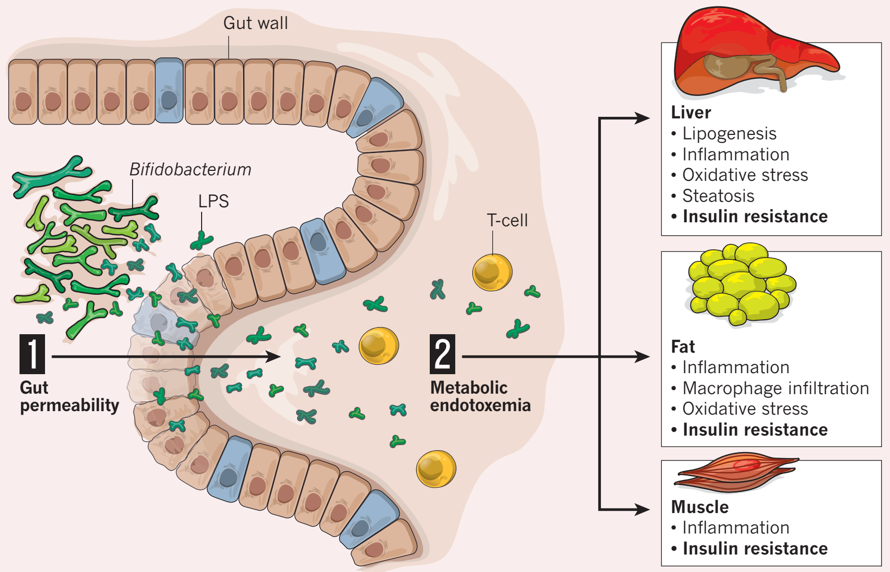
And second, what mechanisms are involved in any metabolic change? The answers to these questions will ultimately inform research on both the prevention and treatment of diabetes.

At the moment, researchers are trying to figure out precisely how the gut microbiome is influencing the metabolism, and thus the development of diabetes, of its human host. Several theories exist. One, for instance, blames the metabolites and other chemicals excreted by the bacteria. Another theory implicates the immune system's reaction to the bacterial cells themselves (see 'Microbial influence').

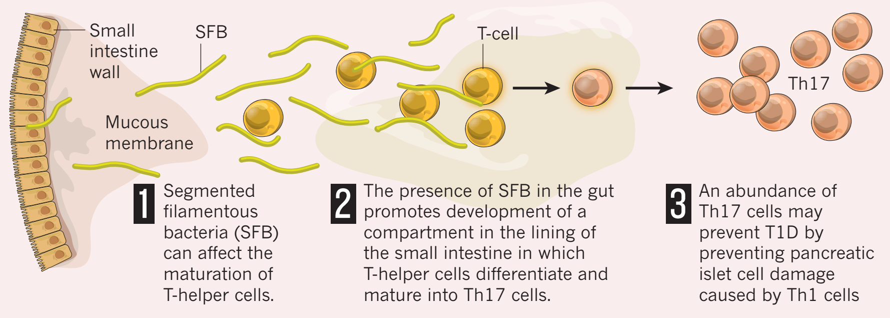
Whatever the mechanism, the bacterial changes that precede insulin resistance can often be attributed to changes in diet. In mice, it takes only one day after switching from a low-fat to high-fat diet for insulin resistance to be detectable[3]. In type 2 diabetes, many researchers believe there is a web of complex interactions between a person's genome and gut flora. Some people are genetically predisposed to have more beneficial bacteria, while others people's guts may be hospitable to pathogenic strains and may be more likely to develop diabetes when they eat high-fat foods. "Your own human nuclear genome controls a considerable part of your individual gut microflora," says Oluf Pedersen, head of diabetes genetics research at the Hagedorn Research Institute in Gentofte, Denmark. "But if your microbiota go off kilter then they can be causative and, at least in rodent models, effect a major change in phenotype." Such phenotypic changes might include weight gain and the development of metabolic syndrome — a precursor to diabetes.

MICROBIAL INFLUENCE

Research by Patrice Cani, at the Université Catholique de Louvain in Brussels, has shown that, in mice, a decrease in the population of bifidobacteria species in the gut causes the tight junctions between the cells of the gut lining to loosen. The loose junctions increase the gut's permeability and allow lipopolysaccharide (LPS) from these microbes to leak through the gut wall. The resulting metabolic endotoxaemia causes a low-grade inflammation and can induce a number of metabolic disorders – including the insulin resistance that characterizes T2D.



Research by Harvard immunologist Diane Mathis suggests that certain bacteria may protect against T1D.



If researchers can figure out which bacterial species of the mammalian gut are beneficial and which are pathogenic, they might be able to nudge the population away from diabetes or even cure it. But with such a vast number of species, many of them never before identified and nearly impossible to culture, developing an extensive profile of the bacteria associated with lean versus insulin-resistant individuals is proving to be monstrously difficult.

Pedersen is tackling that task in his work with the MetaHIT (Metagenomics of the Human Intestinal Tract) consortium, a collaboration of 13 institutions working to understand how genes and intestinal microbiota interact to influence health and disease. As head of MetaHIT's obesity effort, he is sorting people according to their metabolic traits, including insulin resistance, and trying to correlate those to their gut bacteria. In parsing the data, Pedersen and his colleagues are finding that they can sort people into two groups according to the quantity of bacterial genes they have. Roughly one-third of their obese subjects fall into the 'low-gene-count' group. These individuals are more likely to have signs of inflammation, such as high white-blood-cell counts and elevated levels of C-reactive protein. In the general population, about the same fraction of obese people, 30% to 40%, are at risk of developing diabetes. "We seem to have identified a subgroup of obese individuals who have a greater risk of progressing to type 2 diabetes," Pedersen says.

HARNESSING A MICROBE

Establishing that gut flora play a role in causing diabetes is a start, but until scientists can pin responsibility on specific bacterial species or genera, it will be difficult to apply this knowledge to developing diabetes treatments.

One research group took the fecal transplant method that Bäckhed and others had used in mice and adapted them for human testing. Max Nieuwdorp, an endocrinologist at the Academic Medical Center in Amsterdam, the Netherlands, led a team that tested fecal transplants in a trial of 18 men recently diagnosed with metabolic syndrome. Nine men received gut biota from lean donors, while the others had their own microbiota returned to them via a fecal transplant, similar to the procedure used in mice. Initial results provide tantalizing hints that manipulating gut microflora can improve health. After 6 weeks, the men who received transplants from lean donors showed improved insulin sensitivity — an indication that their road to type 2 diabetes had slowed or even halted. One year later, however, the subjects' microbiomes, and their insulin sensitivity, had returned to their original states[4].

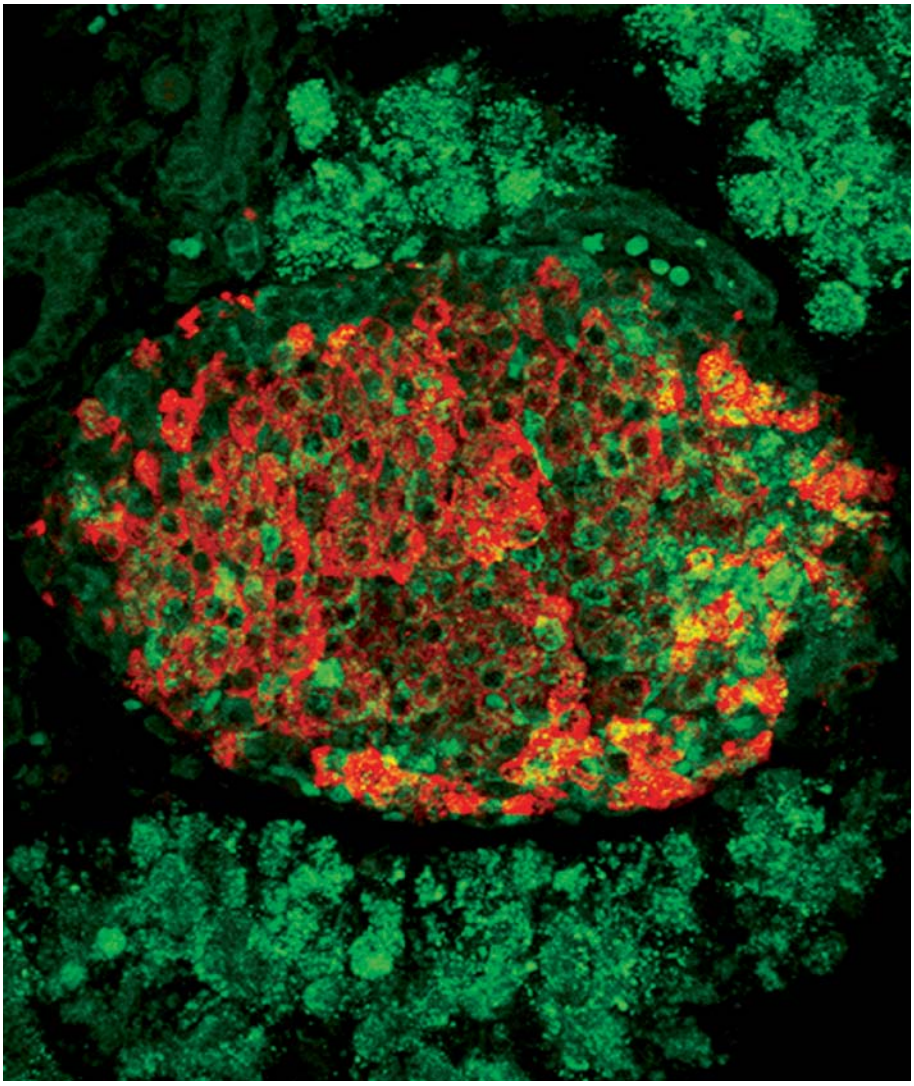
Fecal transplants in their current form aren't a practical cure for diabetes or obesity; there are too many risks, including the transfer of bacterial infections from donor to recipient. But these transplants do confirm the impact of bacterial composition on blood sugar regulation in humans. And if researchers can figure out which bacteria are beneficial, and why, they might be able to develop drugs or bacterial supplements that mimic those effects. "No one knows whether there's a causal relationship between bacteria and diabetes," Nieuwdorp says. "We tried to show it with the fecal transplant, but the only thing we can say is that there seems to be a transmissible trait." Nieuwdorp has already begun a longer trial of 45 people in conjunction with gene-chip testing to discover whether multiple transplants might produce a longer-lasting effect and to identify the bacterial species involved.

If scientists can determine which bacteria are associated with which metabolic profile (lean and insulin sensitive versus overweight and insulin resistant), they might be able to supplement accordingly. Probiotics (live bacteria) and prebiotics (which encourage the growth of beneficial bacteria) could be used to tune a person's microbiome towards greater insulin sensitivity. Antibiotics could be designed to target pathogenic species, or prescribed in conjunction with supplements of beneficial bacteria to prevent irreparable harm. And if researchers can identify the mechanisms of action, they should be able to develop drugs that mimic the chemicals produced by the bacteria found in lean people's guts, or inhibit the metabolites or other molecules that lead to insulin resistance and diabetes.

"We don't think the gut microbes are acting by one mechanism but by a contribution of several," says Bäckhed, now at the University of Gothenberg in Sweden. "We don't know what we do when we change the microbiota yet — we might cure type 2 diabetes and predispose someone to type 1. I wouldn't say that changing the microbiota could cure everybody, but I think that together with lifestyle changes it could help a lot of people." ■

Lauren Gravitz *is a science writer based in Los Angeles, California.*

1. Backhed, F. *et al. Proc. Natl. Acad. Sci. USA* **101**, 15718–15723 (2004).
2. Turnbaugh, P. J. *et al. Nature* **444**, 1027–1031 (2006).
3. Turnbaugh, P. J. *et al. Sci. Transl. Med.* **1**, 6ra14 (2009).
4. Vrieze, A. *et al. Diabetologia* **53**, 606–613 (2010).



The Coxsackie virus has been linked to diabetes, but do viral infections trigger or stave off diabetes?

FLODSTROM, M. *ET AL. NAT. IMMUNOL.* **3**, 373–382 (2002).

PATHOLOGY

Cause and effect

Decades of study into the causes of diabetes have produced no definitive answers.

BY ERIKA JONIETZ

Type 1 and type 2 diabetes have long been viewed as two diseases, the first auto-immune with a large genetic component, the second metabolic, linked to obesity and a sedentary lifestyle. "These are two very different disorders," says C. Ronald Kahn, a senior investigator at Joslin Diabetes Center and professor of medicine at Harvard Medical School in Boston, Massachusetts. "They lead to similar metabolic problems and similar long-term complications, but they have two very different pathogenic routes."

Researchers, however, are showing that each type has more in common with the other than once believed: both involve a faulty immune system and share some mechanisms that ultimately kill the insulin-producing beta cells in the pancreatic islets. Yet in neither type 1 diabetes (T1D) nor type 2 diabetes (T2D) does genetics or behaviour fully explain why some people get the disease and others don't.

Recent findings show that despite some common risk factors, the two are indeed separate conditions. In addition to the many genetic factors involved, scientists have implicated epigenetic and environmental influence in each type of diabetes. Researchers continue to search for certain causes in an effort to prevent both.

ALL IN THE FAMILY?

Type 1 diabetes is an autoimmune disease in which the immune system kills insulin-producing beta cells. It runs in families, the hallmark of any genetic disease. About 60% of the genetic risk comes from a few specific variants in the human leukocyte antigen (HLA) genes. These genes encode the proteins that present antigens to immune cells and are involved in the misguided immune response in T1D.

Better understanding HLA, therefore, could help unravel the origins of T1D. Over the past five years, George Eisenbarth, an endocrinologist at University of Colorado Medical School in Denver, along with immunologist John Kappler at National Jewish Health, has been working out the structure of a three-protein complex he believes is the crux of the disease. The complex consists of an antigen-presenting HLA molecule, the antigen itself (a specific insulin peptide) and a T-cell receptor that recognizes the HLA–antigen combination.

T cells are central to all autoimmune diseases, including type 1 diabetes. Normally, cytotoxic T cells destroy only infected cells; T cells that react to molecules native to the body are eliminated before they mature, thus endowing the immune system with tolerance to 'self'. In type 1 diabetes, however, things go awry: T cells primed to recognize beta cells enter circulation and go on to attack the cells. How these T cells escape destruction and reach maturity isn't clear. A number of factors appear to be involved, including variations in the gene encoding insulin, diet, and the presence or absence of certain bacteria in the gut flora (see 'The critters within', page S12).

Eisenbarth was involved in much of the early work that identified the antigens that prime T cells against beta cells; besides insulin, the major autoantigens are ZnT8, GAD65 and IA-2. "By following the development of antibodies to these four antigens," Eisenbarth says, "we can now predict diabetes." He adds: "Whoever has two of those, they almost all get diabetes."

Further insight into the origin of diabetes could come from a new technology that can track the development of the disease in humans and mice. Researchers at Harvard Medical School and Massachusetts General Hospital in Boston have used magnetic resonance imaging (MRI) of magnetic nanoparticles to visualize insulitis, the inflamed pancreatic tissue that is the earliest clinical manifestation of diabetes. They also used MRI to distinguish at just 6–10 weeks of age which non-obese diabetic (NOD) mice — a model of T1D — will develop full-blown diabetes; mice with the highest pancreatic accumulation of the magnetic nanoparticles, used as a probe, got diabetes.

➲ NATURE.COM
Is it time to reclassify autoimmune disease?
go.nature.com/bkxr2g

"We now have a way to know very early whether [mice] will or won't get diabetes and then compare them at the molecular level," says Diane Mathis, an immunologist at Harvard Medical School. Performing these comparisons has enabled her group to identify several previously unknown molecular and cellular elements associated with a lower chance of the mice developing diabetes.

As not everyone with a genetic susceptibility to T1D actually develops the disease, some sort of trigger might be involved. Evidence suggests that a viral infection — possibly by enteroviruses such as the Coxsackie virus — causes the immune system to misbehave. There are two theories about viral exposure: one suggests that viruses and other microorganisms improve tolerance and may thus protect against T1D. The presence of such pathogens might help overcome a sort of "boredom of the immune system" resulting from fewer childhood infections, says Matthias von Herrath, director of the Type 1 Diabetes Center at La Jolla Institute for Allergy and Immunology in California. The other theory is that a virus somehow exposes antigens on beta cells, causing the immune system to attack them.

To determine whether viruses or anything else in the environment trigger type 1 diabetes, the Diabetes Auto Immunity Study in the Young (DAISY) began in July 1993. DAISY HLA-typed about 30,000 newborns and enrolled children with a parent or sibling with T1D or children in the general population with genetic markers that indicated they were at moderate or high risk for the disease. Researchers collected blood samples and interviewed parents about diet, health and other aspects of their children's lives; as of February 2007, 61 of the children were diagnosed with type 1 diabetes. According to the DAISY organizers, the team identified several autoantigens and genes associated with T1D over 15 years. They also linked diet to the onset or delay of diabetes, and disproved any association between type 1 diabetes and the age of childhood vaccinations. And although a small prospective study found no link between enterovirus infection and T1Ds, the team noted the need for more studies.

IT'S COMPLICATED

The search for the triggers of type 2 diabetes is not any easier. This condition occurs when muscle and fat tissue respond abnormally to insulin, together with a failure of beta cells to compensate by pumping out more insulin. The statistical connection between T2D and a high-calorie diet and sedentary lifestyle is well established, but researchers still debate how — or whether —these factors cause the initial resistance to insulin. After all, 75-80% of obese people never develop type 2 diabetes. Moreover, as with type 1 diabetes, type 2 diabetes seems to run in families. Together these data suggest genetic elements.

The data from genome-wide association studies (GWAS) are far from clear, however. Studies so far have identified more than 40 genes associated with T2D, most of them having to do with beta-cell function[1]. But added together, they account for only about 10% of the apparent genetic causes. To find the missing heritability, biochemist Alan Attie at the University of Wisconsin-Madison has crossbred two strains of mice used as models — one obese but non-diabetic, the other obese and prone to diabetes — to hunt down genes linked to intermediate processes involved in diabetes, such as those that govern beta-cell regeneration, insulin degradation and insulin secretion.

MRI of a mouse pancreas (colour) tracks disease progression at the cellular level.

G WOJTKIEWICZ AND R. WEISSLEDER

Some genes implicated by GWAS are expressed only in adipocytes (fat cells), which might help explain how overeating can lead to diabetes. Adipose tissue stores excess lipids, which are otherwise toxic to the body. When fat cells malfunction and aren't able to store away the extra lipids generated by overeating, lipids begin to accumulate in muscle tissue and in the liver. Philipp Scherer, a diabetes researcher at the University of Texas Southwestern Medical Center in Dallas, believes this aberrant build-up triggers insulin resistance. When adipose tissue expands in a healthy way there is no insulin resistance — which for Scherer explains why some obese people never develop type 2 diabetes.

"Researchers in the field are confused and have different opinions."

Another consequence of abnormal adipose growth is inflammation. Expanding fat mass produces proteins called cytokines and other substances that promote inflammation and recruit macrophages (killer immune cells). As macrophages accumulate in adipose tissue, they change and secrete even more cytokines and other inflammatory factors into the bloodstream. This promotes inflammation in other tissues, including pancreatic islets. Researchers agree almost unanimously that insulitis plays a role in type 2 diabetes; the nature of that role, however, is still a matter of debate.

While Scherer sees the inflammation as a result of insulin resistance, other biologists believe inflammation is a primary cause of diabetes. Steven Shoelson, a doctor and structural biologist at Joslin Diabetes Center and Harvard Medical School, sees things the latter way: he believes that cytokines released in response to metabolic stress may directly lead to insulin resistance. Shoelson is involved in trials to assess whether salsalate, a non-steroidal anti-inflammatory drug, can lower levels of sugar and lipids in the blood of patients with T2D, and plans to present the results of the latest large-scale trial of salsalate at the American Diabetes Association meeting in June 2012 in Philadelphia, Pennsylvania.

But even genetics, diet and activity levels combined don't completely explain the origins of type 2 diabetes. Other factors that might contribute include environment toxins and the gut microbiome. Another influence may be maternal diet: research in both mice and humans has shown that maternal caloric restriction during gestation increases the risk of T2D in offspring[2]. The mechanism might involve strong epigenetic programming, Kahn says. Rat and human studies, for example, show that poor diet during pregnancy may affect the expression of genes that influence fetal fat-cell development, making it harder for adipocytes to effectively store excess lipids.

In a novel effort to identify specific environmental factors associated with T2D, Atul Butte, a paediatric endocrinologist and medical informaticist at Stanford University in California, created an environmental-wide association study analogous to GWAS[3]. A pilot study found significant links between T2D and the pesticide derivative heptachlor epoxide, vitamin E and polychlorinated biphenyls (PCBs).

Despite experts' increasing knowledge about both types of diabetes, much about the pathologies of both diseases and virtually everything about their aetiologies remains a mystery. Most researchers acknowledge that it's unlikely there is a single trigger; some even suggest that different genes and environmental factors may lead to disease processes that differ from person to person. "Researchers in the field are confused and have different opinions," says Shoelson. Whether scientists ultimately find the factors that cause diabetes or not, Scherer agrees, "the bottom line is, it's complicated". ■

Erika Jonietz *is a science writer based in Austin, Texas.*

1. Fu, W. *et al. Nat. Immunol.* **13,** 361–368 (2011).
2. Herder, C. & Roden, M. *Eur. J. Clin. Invest.* **41,** 679–692 (2011).
3. Patel, C. J., Bhattacharya, J. & Butt, A. J. *PLoS ONE* **5,** e10746 (2010).

PERSPECTIVE

Testing failures

Promising drugs to treat diabetes stumble in the latter stages of clinical testing. **Thomas Mandrup-Poulsen** explains why — and how to fix it.

The development of certain diabetes drugs keep hitting a snag — phase III clinical trials. This final stage of clinical testing is designed to test the efficacy and safety of treatments in 300 to 1,000 or more patients to ensure that the results from earlier trial phases can be applied to a more general population. Recently, a striking pattern has emerged: trials are failing to confirm encouraging results obtained in earlier trials. In particular, recent phase II studies of short courses of immunomodulatory biologics have provided proof-of-principle that this strategy can at least transiently improve glycaemia, insulin sensitivity or beta-cell function in people with type 1 and type 2 diabetes (T1D and T2D). Four to six infusions of antibodies against the common T-cell surface marker CD3 (ref. 1) or the B-cell surface antigen CD20 (ref. 2) — both central determinants of adaptive immunity — preserved beta-cell function and/or reduced insulin needs after 12–18 months in groups of 80–90 patients with recent-onset T1D. In 70 long-term patients with T2D, a blocker of the receptor binding interleukin-1 (IL-1), the primary inflammatory mediator of innate immunity, resulted in an improvement in beta-cell function — an effect that lasted throughout the 39-week follow-up[3,4]. These trials created optimism for the success of these agents in later phase trials.

Disappointingly, the larger trials of these drugs have failed to meet their primary clinical endpoints — the measure of a trial's success. Careful analysis has pointed to important differences in the design of the phase II and III trials. In the case of anti-CD3 antibody, the Protégé phase III study of more than 500 patients with new-onset T1D used a dose regimen different from that of the companion phase II study[5]; moreover, this study, conducted by MacroGenics, selected glycaemia and insulin needs as primary endpoints, instead of beta-cell function (the phase II endpoint). Another anti-CD3 study, Defend-1, conducted by GlaxoSmithKline, used beta-cell function as an endpoint. Because the full study results have not been published, we do not know such important details as whether beta-cell function was measured during fasting or after meal stimulation as generally recommended. Furthermore, the study used a 15-fold lower dose than that effective in phase II.

Similarly, a large phase IIb trial of IL-1 blockade conducted by XOMA, a firm in Berkeley, California, and not yet published, enrolled more than 400 patients with T2D. The trial subjects were on average 6 years post-diagnosis, and were maintaining a baseline glycaemia of 7.8% on a single oral antidiabetic (less than 6% is considered a healthy level). In contrast, patients in the phase II trial were taking a combination of oral antidiabetics and insulin, and had a mean disease duration of 11 years and baseline glycaemia of 8.5% (ref. 3). So the subjects in the larger trial had a shorter disease duration and better glucose control than those enrolled in the proof-of-principle study.

THE DEVELOPMENT OF CERTAIN DRUGS KEEPS HITTING A SNAG — PHASE III TRIALS RARELY CONFIRM ENCOURAGING RESULTS

This experience prompts the question: were the right drugs tested at a wrong dose or in the wrong patients? Post-hoc analysis of the Protégé study did find significantly improved glycaemia and reduced insulin needs in the cohort receiving the highest dose[5], suggesting that patients with new-onset T1D are highly sensitive to the dosing of anti-CD3 antibody. This subgroup analysis also suggests that insufficient doses might account for the failure of the Defend-1 trial. Finally, there is preclinical evidence that IL-1 blockade is more effective at preserving insulin secretion when the glucose drive is high.

Changes in study rationale, dosage, patient selection and clinical endpoints may compromise the ability to confirm phase II findings in larger trials. The implications for drug development are clear, and organizers of new trials would be well advised to consider the following:

1. Recognizing that certain therapies may only be effective in subsets of patients, phase III trials should use entry criteria and endpoints as close as possible to those used in phase II, and generalization of the outcomes to the prescribed patient population could then be broadened by less restrictive exclusion criteria.
2. Phase III trials should include the doses and dosing regimens effective in phase II.
3. Negative results should be published to allow learning from failure.
4. Collaboration between academia and industry should be promoted to ensure that trial designs are based on the strongest experimental and empirical evidence.

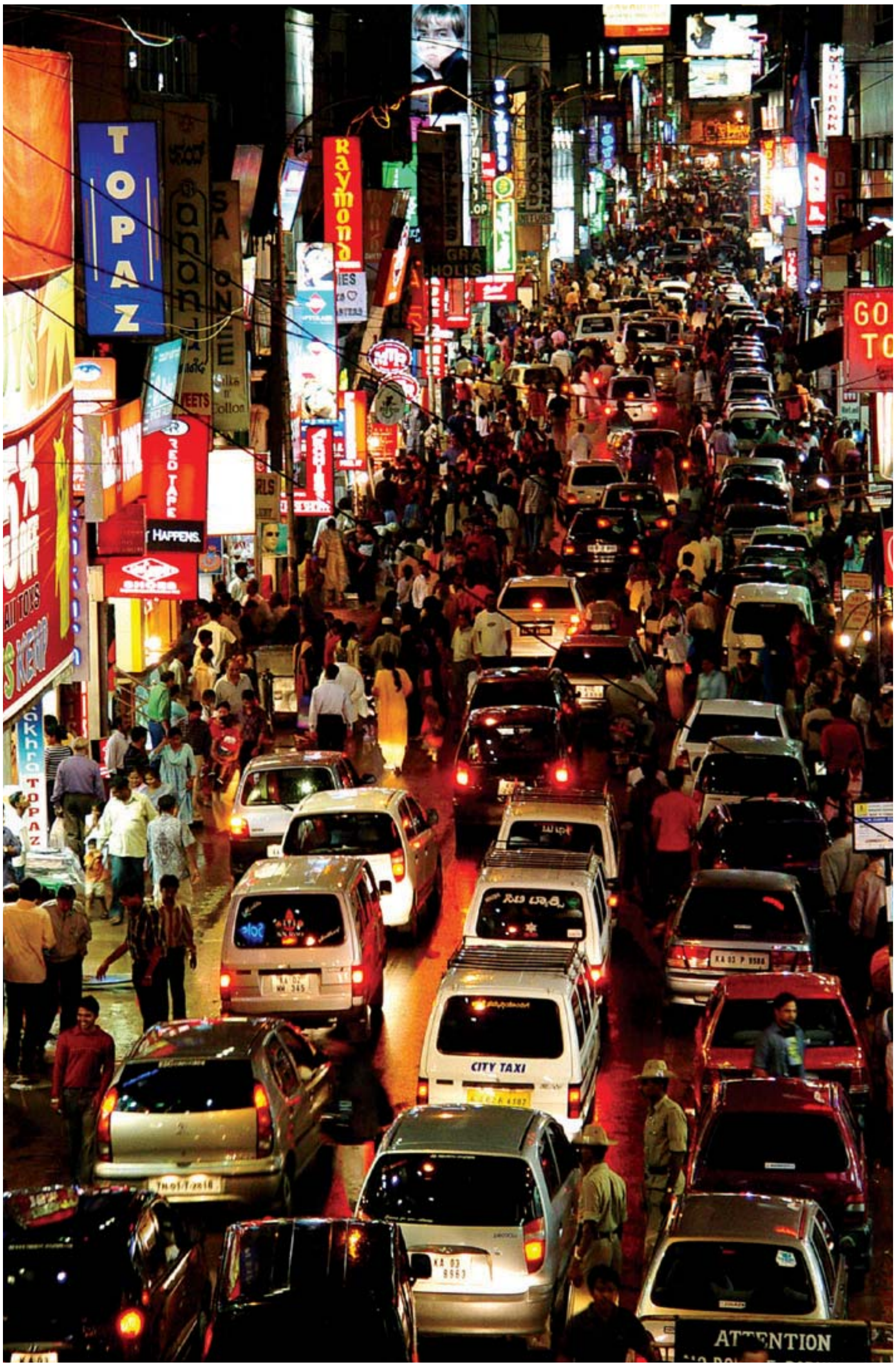
These may be more general implications for developers of drugs to treat chronic degenerative diseases, for which current clinical classifications are too crude to discriminate between aetiologically and pathogenetically different populations of patients that may require different management.

Industry and academia are in this boat together. In these times of financial constraints, with growing rates of attrition in industry and funding sources drying up in academia, there has never been a greater need for trustworthy public–private partnerships. ■

Acknowledgements
Marc Y. Donath and Charles A. Dinarello are thanked for comments.

Thomas Mandrup-Poulsen *is an endocrinologist at the University of Copenhagen's Institute of Biomedical Sciences, Denmark, and at the Karolinska Institute in Stockholm, Sweden.*
email: tmpo@sund.ku.dk

1. Keymeulen, B. *et al. N. Engl. J. Med.* **352,** 2598–2608 (2005).
2. Pescovitz, M. D. *et al. N. Engl. J. Med.* **361,** 2143–2152 (2009).
3. Larsen, C. M. *et al. N. Engl. J. Med.* **356,** 1517–1526 (2007).
4. Larsen, C. M. *et al. Diabetes Care* **32,** 1663–1668 (2009).
5. Sherry, N. *et al. Lancet* **378,** 487–497 (2011).



JAGADEESH N.V/REUTERS/CORBIS

PUBLIC HEALTH

India's diabetes time bomb

Epigenetics and lifestyle are conspiring to inflict a massive epidemic of type 2 diabetes in the subcontinent.

BY PRIYA SHETTY

Mumbai's Linking road, a congested artery at the heart of the city's eternal traffic jam, offers a disturbing snapshot of the way that growing wealth is compromising India's health. Roadside carts selling traditional fried sweets and samosas jostle with fast-food joints selling burgers and fries, and shopping malls are full of shops selling myriad labour-saving appliances.

India's embrace of the worst of both Eastern and Western ways is sending lifestyle illnesses such as obesity and diabetes skyrocketing. In 2011, India had 62.4 million people with type 2 diabetes, compared with 50.8 million the previous year, according to the International Diabetes Federation (IDF) and the Madras Diabetes Research Foundation. As the economy started growing, so did the incidence of diabetes. The nationwide prevalence of diabetes in India now tops 9%, and is as high as 20% in the relatively prosperous southern cities. By 2030, the IDF predicts, India will have 100 million people with diabetes.

Health experts are alarmed because, although the onset of type 2 diabetes tends to affect people in the West in their 40s and 50s, the disease strikes Indians much younger. Indians as young as 25 are being diagnosed with the disease, a trend that threatens to seriously hamper the country's economic development.

The rise of type 2 diabetes in India's cities was to some extent expected. And in fact, until the 1980s, the urban prevalence of diabetes was at least double the rural prevalence. But the recent surge in diabetes has spilled out of the cities into the countryside. The spike in rural areas has been shocking, says Nikhil Tandon, an endocrinologist at the All India Institute of Medical Sciences in New Delhi (see 'India's diabetes boom'). "Villages in wealthier southern states like Tamil Nadu and Kerala are seeing prevalence hit double digits, which is enormous," he says. "If it was confined to affluent India, you could still put a lid on it, but now it's rising quickly all over the country."

THRIFTY GENES

Health experts in countries like the United States have for years been lamenting the trend towards overeating and lack of physical exercise, and the resulting rise in obesity, diabetes and heart disease. Indians seem to be even more vulnerable to these lifestyle changes. The culprit may be what is called the 'thrifty genotype', whereby millennia of evolution have shaped the genetic profile to cope with hardship. According to this theory, some of the world's populations, including Indians, are genetically adapted to an environment in which calories are scarce. As a result, their bodies can't cope in times of over-indulgence, and it takes only a small increase in daily calories (or a small drop in calorie expenditure) for their metabolism to tip over into diabetes.

➲ NATURE.COM
visit *Nature India* for latest and best Indian research:
go.nature.com/znowpk

But the idea of a thrifty genotype doesn't fully explain the prevalence of type 2 diabetes among certain populations, says Antonio Gonzalez-Bulnes, a geneticist at the National Institute for Agriculture and Food Research and Technology (INIA), Spain. He points out that although this genotype has been identified in rat models and in particular ethnicities around the world, such as South Pacific islanders, these groups haven't always had higher rates of metabolic disease. C. Ronald Kahn, director of the Joslin Diabetes Center in Boston, Massachusetts, emphasizes that "we still need to regard it as a hypothesis, since no genes have been specifically identified that contribute to the phenotype".

Instead, researchers like Gonzales-Bulnes and Tandon are interested in a related idea of a 'thrifty phenotype': that being deprived of nutrients in the womb, but then exposed to a high-calorie and low-exercise life, leads to a person to develop diabetes. The supposed mechanism is epigenetic; the fetal environment triggers changes in DNA methylation, which is responsible for switching genes on or off. The environment *in utero* thereby affects the expression of genes that code for enzymes that regulate blood sugar or tell our brains when we have eaten enough. "The mother's nutrition, or even her smoking or alcohol consumption, can change the way the baby's genes react to the environment: a poor or excessive diet and sedentary lifestyle," says Paul Zimmet, head of international research at the Baker IDI Heart and Diabetes Institute in Melbourne, Australia, and one of the first to predict the Indian diabetes epidemic.

Several epigenetic studies back up the idea that the *in utero* environment has a life-long influence on health. In one study of the Dutch Hunger Winter of 1944, in which thousands of people starved during a German blockade, children born to women who were pregnant during the famine were far more likely to develop obesity or diabetes; this finding was backed up by studies of children born during the Chinese famine of 1959–61.

IN THE WOMB

The problem of the thrifty phenotype begins before the child is born. A fetus growing in a malnourished mother will need to grab all the glucose it can for its development. It does this by making its muscles resistant to insulin; since insulin is responsible for allowing fat and muscles to store sugar, insulin resistance forces the sugar to circulate in the blood instead. But when food is freely available, this inability to store glucose can send blood sugar levels soaring and trigger the onset of type 2 diabetes.

The maternal link may help explain why the diabetes epidemic is being seen all over India, in both rural and urban areas, says Caroline Fall, a paediatric epidemiologist at the University of Southampton, UK. India already has a problem with babies being born underweight — 40% of 20 million babies born weighing less than 2.5 kilograms in the developing world are born in India. Fall points out that low birthweight (a marker of poor maternal and fetal nutrition)

AP PHOTO/RAJESH KUMAR SINGH

Doctors and students "walk together to keep diabetes away" at a rally on World Diabetes.

does not differ that much between cities and villages in India. "You don't need to have severe maternal malnutrition to produce the problem of obesity and diabetes in later life," Fall says.

If poor maternal nutrition could cause diabetes, might improving it prevent the disease? Fall is investigating the effect of micronutrients such as folate and vitamin B12 in pregnancy on the child's development of diabetes. This link has been proven in animal models, says Fall. She is now trying to see if there is a similar effect in humans, through a study of maternal nutrition in 5,000 women living in a Mumbai slum.

Emerging data lends further support to the prenatal nutrition link. Sanjay Kinra, a chronic-disease epidemiologist at the London School of Hygiene and Tropical Medicine, found that children whose mothers took nutritional supplements during pregnancy had lower insulin resistance[2]. Similar results have been found in the Gambia, says Fall. And studies by Tandon and others of the New Delhi Birth Cohort, which is following people born between 1969 and 1972, reinforce the link. According to Tandon, "those who were born small relative to their peers and then gained weight rapidly —not necessarily becoming obese — are the ones who later in life had the highest risk of developing metabolic diseases[3]".

> *"The question is whether the Indian government is prepared to put in the resources."*

Tandon's findings tie in with previous observations that Indians don't need to be as overweight as people of other ethnicities to develop diabetes. The reason lies in Indians' natural body composition, says Fall. Pound for pound, she says, "Indians have less lean mass, more body fat, and more central fat than a white Caucasian. All of these very much increase the risk of diabetes." This difference in body type affects standard measurements such as body mass index (BMI): according to Fall, an Indian with a BMI of 23 has the same amount of body fat as a British Caucasian person with a BMI of 25. Thus, the BMI threshold that serves as a warning sign for developing diseases like type 2 diabetes is much lower in Indians.

Moreover, says Fall, the Indian susceptibility starts before birth. Even without poor maternal nutrition, "a lower muscle growth but higher fat growth *in utero* makes babies more vulnerable," she says. "This is why we think diabetes hits earlier in India — they are more vulnerable from the start. If babies were well nourished in the womb, it might mean that they were not so biologically susceptible to changes in diet and lifestyle, and therefore more immune to diseases like diabetes."

In addition to improving maternal nutrition, Fall wants to see routine screening for gestational diabetes (high blood glucose in pregnant women) because the condition is known to prime the child to be insulin-resistant and significantly increases the chance that the child will develop diabetes later in life. She points out that gestational diabetes is 5 or 10 times more common in Indian cities than in the United Kingdom.

Tandon points out that the focus on mothers and babies has a corollary: it means that "the problem cannot just be solved by taking 30- or 40-somethings and getting them to exercise", he says. "We've missed the boat if we do that."

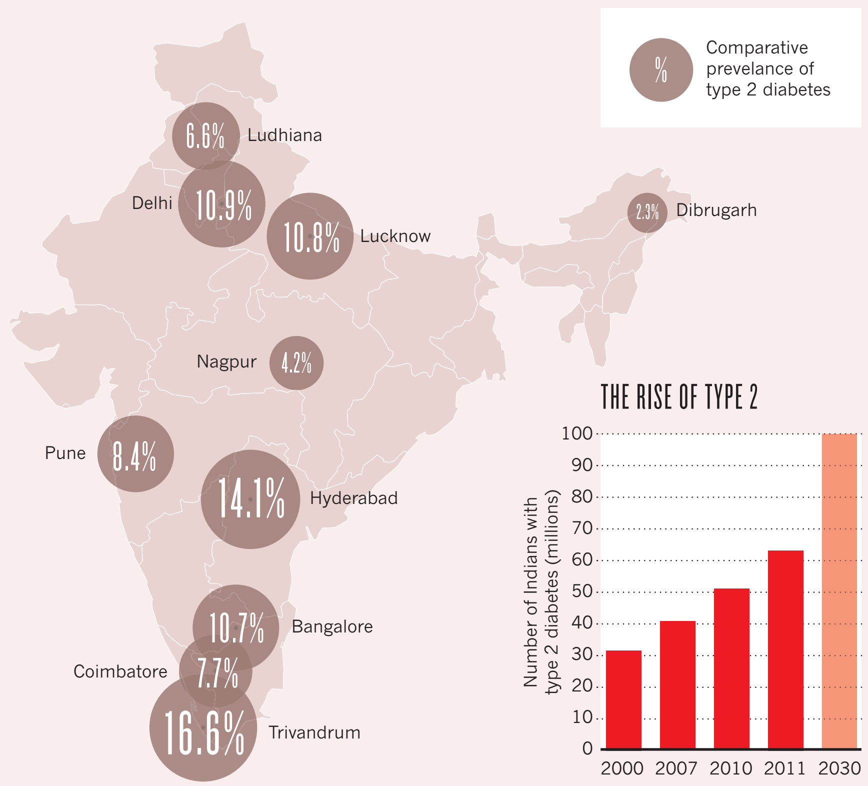
TOO LITTLE ACTION

Screening for gestational diabetes could be implemented as part of a national diabetes prevention plan in India, though that is still being developed slowly. Preventing diabetes should be a high priority for India but there is little evidence that any major initiatives are underway in that direction," says Zimmet. He adds that while "diabetes is now regarded as a very serious problem by the Indian government, the question is whether they are prepared to put in the resources that are needed to turn around the epidemic".

The early signs are that it has at least set the wheels in motion. In its latest 5-year plan, the Indian government has dedicated a significant chunk of funding for non-communicable

INDIA'S DIABETES BOOM

The Western diet and lifestyle that have accompanied India's growing prosperity has brought an alarming rise in cases of type 2 diabetes. Nationwide, prevalence of T2D is more than 9%. The epidemic is not surprising in urban areas. However, the disease is now also becoming common in rural villages, especially in wealthy southern states.



IDF DIABETES ATLAS, 4TH ED. INTERNATIONAL DIABETS FEDERATION (2009). REDDY, K. S. BULL WORLD HEALTH ORGAN. 84, 461–469 (2006).

diseases such as diabetes. On chronic diseases overall, it will spend 580 billion rupees (US$11.6 billion) — six times what was allocated in the previous 5-year plan[4]. Progress is slow, however. India's National Programme for Prevention and Control of Diabetes, Cardiovascular Diseases and Stroke (NPCDS), launched in 2008, has made little headway in either strengthening infrastructure or implementing prevention plans, other than developing a website (healthy-india.org) to educate people about risk factors for chronic diseases such as diabetes. (India's Ministry of Health did not respond to *Nature Outlook*'s requests for comment on the diabetes epidemic.)

> ***"Unless we streamline patient flow, it will be very difficult to handle the growing number of people with chronic diseases."***

India's national programme on diabetes might be at a nascent stage, says Tandon, but he nevertheless finds it "reassuring" that in 100 districts, the government will be targeting programmes at schoolchildren to help reduce diabetes and other health risks in later life. The department of health is also rolling out a much-needed nationwide prevalence study, he says, which should help provide a cohesive picture — most data currently available are from fragmented regional studies, predominantly in cities.

One big barrier to improving diabetes care, says Tandon, is India's chaotic patient referral system. "The lack of systemic processing of patients has been a bugbear of the Indian healthcare system," he says. Patients aren't first seen by primary care workers, and then referred to secondary or tertiary care. Because 80% of healthcare in India is private, many people bypass general practitioners and go straight to the specialists, a habit that overburdens their clinics. "Unless we streamline patient flow in the future, it will be very difficult to handle the growing numbers of people with chronic diseases," says Tandon.

Tackling most epidemics starts with screening, but this is difficult given India's ailing healthcare system and the inability of most people to afford glucose tests. The World Health Organization has recommended HbA1c as a proxy for blood glucose level, as the test is cheaper and quicker than a glucose-tolerance test. Fall points out that this method tends to create a lot of false alarms in India because of the country's unusually high prevalence of iron-deficiency anaemia, a condition that elevates HbA1c levels.

Although better screening would be desirable, Zimmet argues that the top priority needs to be prevention. He points out that India won't have the resources to treat growing numbers. "Rather, India needs to look to the future, and that may be 20 or 30 years down the track, to reducing the burden. Attention to maternal and child health may be an important way of eventually stemming the epidemic."

For India to effectively fight the onslaught of diabetes will require more than government programmes, however. Indian society's nonchalant attitude towards the disease must change as well. "Almost 50% don't follow any diet and exercise regime despite our advice, and 25% of the rest will follow it initially but then abandon it," says Anoop Misra, head of diabetes and metabolic diseases at Fortis Hospital in New Delhi. People believe that since they don't have symptoms from their high-sugar levels, "they don't need to worry about something that won't harm them until a decade later", says Misra.

Tandon agrees, saying that studies of diabetes awareness, especially in urban areas, have shown "pathetic" results[5]. "If you're only worried about peeing a bit more in the night, without realizing that this is a disease that could blind you, knock your kidneys out or give you a heart attack, you won't worry too much."

Ignorance about diabetes can be lethal when combined with the fact that many patients first seek out alternatives such as homeopathy or the traditional Indian medicine known as ayurveda, says Misra. He estimates that about 10–15% of his patients first tried traditional medicines — a detour that can delay their treatment through conventional medicine by up to a year.

The Indian government has been trying to raise awareness with television advertisements and posters about diabetes in doctor's clinics and hospitals. Tandon welcomes these efforts. But considering the momentous cultural and political shifts required, he's cautious about how quickly change will come. "It's still an incredibly long haul."

The middle-aged, middle-class people passing through Mumbai's shopping mecca consider themselves to be part of the new, prosperous India. But as the epidemic of type 2 diabetes takes hold, many of them face a chronic condition — and a fate that could signal a warning for other parts of the world that are starting to enjoy abundant food and freedom from labour. ■

Priya Shetty *is a science writer based in London.*

1. Hales, C. N. & Barker, D. J. *Diabetologia* **35,** 595 (1992).
2. Kinra, S. *et al. Brit. Med. Journ.* **337,** 605 (2008).
3. Sachdev, H. P. *et al. Arch. Dis Child.* **94,** 768–774 (2009).
4. http://articles.timesofindia.indiatimes.com/2012-01-17/india/30634918_1_chronic-ncds-diseases-chronic-kidney
5. Mohan, D. *et al. J. Assoc. Physicians India* **53,** 283–287 (2005).

TIM CLAYTON/CORBIS

PREVENTION

Nipped in the bud

While type 1 diabetes might be promising ground for a vaccine, the most effective way to avoid type 2 remains good old-fashioned diet and exercise.

BY SCOTT P. EDWARDS

They share a name, are characterized by elevated blood glucose levels, and carry potentially devastating complications if left uncontrolled. But beyond that, type 1 and type 2 diabetes could not be more dissimilar, says diabetes researcher John Buse, director of the Diabetes Care Center at the University of North Carolina. And perhaps the biggest difference of all, he says, is in their preventability. "With type 1 diabetes, there's an immune process at play," says Buse, "and we don't have effective ways to prevent it. Type 2 diabetes (T2D) has classic risk factors that can be modified to either delay the onset of the disease or prevent it completely."

In the United States, 90–95% of the 17.9 million people diagnosed with diabetes have T2D. Before developing the disease, most people almost always have a related condition, called prediabetes, in which their blood has higher concentrations of glucose than is considered normal, but not high enough to signify diabetes. The American Diabetes Association estimates that 79 million people in the United States have prediabetes and thus are at high risk for developing T2D. Similar to T2D, one of the top risk factors for prediabetes is excess body weight, especially when fat is carried around the abdomen, indicative of physical inactivity and overconsumption. In general, says Buse, people with prediabetes have several problems, including insulin resistance and impaired insulin secretion, so "the train has left the station, and unless these people make changes to reduce their risk, many will keep hurtling down the tracks" towards T2D.

CHANGING LIFESTYLES

Scientists and clinicians have long known that people who change their lifestyle, such as by eating a healthier diet, losing weight and exercising more, lower their chances of developing T2D. Ten years ago, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), part of the US National Institutes of Health, confirmed this point of view when it released the findings of the Diabetes Prevention Program (DPP). This multi-centre clinical research study aimed to determine if lifestyle modifications or treatment with an oral diabetes drug (metformin) could prevent or delay the onset of T2D. The answer was an unequivocal yes to both: according to the NIDDK, "millions of high-risk people can avoid developing type 2 diabetes" by losing 7% of their body weight — and maintaining that loss — by eating less fat and fewer calories, and by exercising for at least 150 minutes per week. The study found that diet and exercise interventions reduced the risk of a person developing T2D by 58%. Lifestyle changes were shown to be particularly effective in older people; those 60 years and older reduced their risk by 71%. The study also found that metformin, an oral drug widely used for the treatment of T2D, can help forestall onset of the disease and reduced risk by 31%, most effectively in young, overweight people.

More recently, researchers in China studied the long-term effects of intensive lifestyle

interventions on the incidence of T2D in those with impaired glucose tolerance, another precursor of T2D (ref. 1). Six years of consuming a diet rich in vegetables and low in alcohol and simple sugar, as well as 20 minutes of moderate exercise, delayed the onset of T2D for as long as 14 years, although the majority of participants still developed T2D.

STOPPING THE UNSTOPPABLE

The evidence is clear on T2D prevention, but the picture with type 1 diabetes (T1D) is opaque. "At this point," says Jay Skyler of the Diabetes Research Institute, part of the University of Miami, Florida, "there is no known way to prevent or lower the risk of developing type 1 diabetes." Studies in animals in which oral insulin is given before diabetes develops, however, have been shown to delay the onset of disease. Given that the immune system reacts differently to drugs whether given in oral form or injected subcutaneously (the normal route of delivery for insulin), Skyler and other scientists suspect that cells in the digestive tract might play an important role in delaying or mitigating the immune response. "When an antigen, in this case insulin, is presented across a mucosal barrier like the digestive tract," Skyler says, "the immune system forms protective immunity [against disease] rather than destructive immunity [such as an autoimmune disorder]."

This hypothesis is supported by recent human research at the University of South Florida in Tampa. Treatment with oral insulin was found to delay the onset of disease in high-risk relatives of people with T1D (ref. 2). The study also showed that oral insulin could postpone the onset of T1D in those with insulin autoantibodies for as long as four years, but once treatment ceased, patients developed T1D at the same rate as those taking a placebo. The task now, says Skyler, is to determine the mechanisms involved. Once the process is better understood, clinicians can assess the proper dosage at which oral insulin becomes an effective preventive therapy.

The South Florida study was a follow-up to the Diabetes Prevention Trial-Type 1, or DPT-1, which ran from 1994 to 2003 under Skyler's direction. The original DPT-1 study helped to establish how to predict T1D risk and provided insight into the immune events that lead to the development of the disease, but it brought scientists no closer to a prevention strategy. One result, for example, was that low-dose insulin injections do not prevent T1D in people who have a high risk of developing the disease within five years.

Because T1D is an autoimmune disease, one option to prevent disease could be vaccines. Many people with T1D have antibodies to an enzyme found in the brain and pancreas called glutamic acid decarboxylase (GAD). Among other things, GAD is an autoantigen, which activates a subset of T cells that react to GAD as an ally rather than an enemy. In T1D, these friendly T cells can quell the attack against the beta cells. However, results of a recent trial using a GAD-based vaccine were less than promising: treatment with an alum-formulated GAD vaccine, which contains an adjuvant to boost the body's response to antigens, did not preserve beta-cell function in patients with T1D (ref. 3). Nevertheless, a phase II trial of another GAD vaccine, called Diamyd (produced by a company of the same name in Stockholm, Sweden), is set to evaluate whether preventive treatment with the vaccine can delay or halt the progression of T1D in children with a high risk of developing T1D.

Preliminary results of another diabetes vaccine study show that it might be possible to reverse T1D using an inexpensive and long-used tuberculosis vaccine. In animal studies, the Bacillus Calmette-Guérin (BCG) vaccine prevented T cells from destroying insulin-producing cells and allowed the pancreas to once again ramp up insulin production. BCG is an attractive vaccine candidate because it raises levels of tumour-necrosis factor (TNF), an immune protein that can suppress the attack on the pancreas. Autoimmunity specialist Denise Faustman of Massachusetts General Hospital in Boston reported at a 2011 American Diabetes Association meeting in San Diego, California, that low doses of the BCG vaccine temporarily increased insulin production in patients who have had T1D for more than 20 years. In a recent study, Faustman showed that the pancreas actually slowly declines over decades, rather than weeks or months[4]. "This is the first clue we've got about how to kill bad T cells in humans with long-term disease and in which the pancreas kicked in to produce insulin," Faustman says. In a phase II trial, which has begun pre-screening subjects, Faustman and her colleagues will see just how far they can encourage the pancreas to re-establish insulin secretion.

"Millions of high-risk people can avoid developing type 2 diabetes" by losing 7% of their body weight.

If Faustman's vaccine can truly restore insulin production, it could in theory serve as the basis for a prophylactic T1D vaccine. "At all stages of diabetes," she says, "we need to slow or prevent the deterioration of the pancreas' ability to produce insulin or, even better yet, restore insulin secretion to higher levels as we have started to do." ■

Scott P. Edwards *is a freelance science writer based in Holliston, Massachusetts.*

1. Li, G. *et al. Lancet* **371,** 1783–1789 (2008).
2. Vehik, K. *et al. Diabetes Care* **34,** 1585–1590 (2011).
3. Ludvigsson, J. *et al. N. Engl. J. Med.* **366,** 433–442 (2012).
4. Wang, L., Lovejoy, N. F. & Faustman, D. L. *Diabetes Care* **35,** 465–4670 (2012).
5. Mingrone, G. *et al. N. Engl. J. Med.* **366,** 1577–1585 (2012).
6. Schauer, P. R. *et al. N. Engl. J. Med.* **366,** 1567–1576 (2012).

THE SURGICAL SOLUTION

Shrinking the stomach

DR. BARRY SLAVEN, VISUALS UNLIMITED /SCIENCE PHOTO LIBRARY

Is weight-loss surgery the next step in diabetes prevention? Two new studies reported in the *New England Journal of Medicine* suggest it might be in obese people with type 2 diabetes (T2D).

A team at the Catholic University in Rome compared two weight-loss surgery procedures with typical diabetes treatment[5]. After two years, 95% of patients receiving a biliopancreatic diversion and 75% of those receiving a Roux-en-Y gastric bypass were in disease remission with normal blood glucose levels. Both procedures shrink the stomach to the size of a chicken's egg and bypass portions of the small intestine, restricting food absorption; in addition, biliopancreatic diversion removes part of the stomach. None of the patients in a group receiving only medical treatment — consisting of oral diabetes drugs or insulin, modified diet, and increased physical activity — went into remission.

The second study, conducted at the Cleveland Clinic in Ohio, compared patients who had either gastric bypass or sleeve gastrectomy, which cuts the stomach to the size of a banana, to those who received intensive treatment with diet, exercise and medication[6]. One year after surgery, 42% of the gastric-bypass patients and 37% of those having sleeve surgery were in remission, compared with only 12% of the patients treated but not operated on.

Previous observational studies have shown a connection between weight-loss surgery and reduced incidence of T2D, but there was until now no solid evidence making the connection. "This is an important result," says John Buse, a diabetes researcher at the University of North Carolina. "As a randomized study, it is the first proof of the clinical observations that had previously been made." — *S. P. E.*

PLANETARY SCIENCE

Martian sand blowing in the wind

High-resolution spacecraft images show surprisingly large rates of sand transport on Mars. This finding suggests that the planet's surface is a more active environment than previously thought. SEE LETTER P.339

JASPER KOK

The ubiquitous sand dunes on Mars appeared almost motionless when the Viking and subsequent space missions examined them[1]. This made sense. After all, the atmosphere of Mars is about one hundred times less dense than that of Earth, so putting sand into motion on Mars requires rare hurricane-like wind speeds. Consequently, some studies have hypothesized that many Martian dune fields were formed in a previous climate, in which the planet had a thicker atmosphere than it has now[2]. But studies in recent years[3–5] have found surprising signs of activity on the surface of Mars, with movement of sand dunes and ripples detected across the desert planet. On page 339 of this issue, Bridges *et al.*[6] present the first extraterrestrial measurements of sand transport, in a Martian dune field called Nili Patera. Remarkably, the authors report that Mars's thin atmosphere blows sand in this dune field at rates not much lower than Earth's much thicker atmosphere does on terrestrial dunes*.

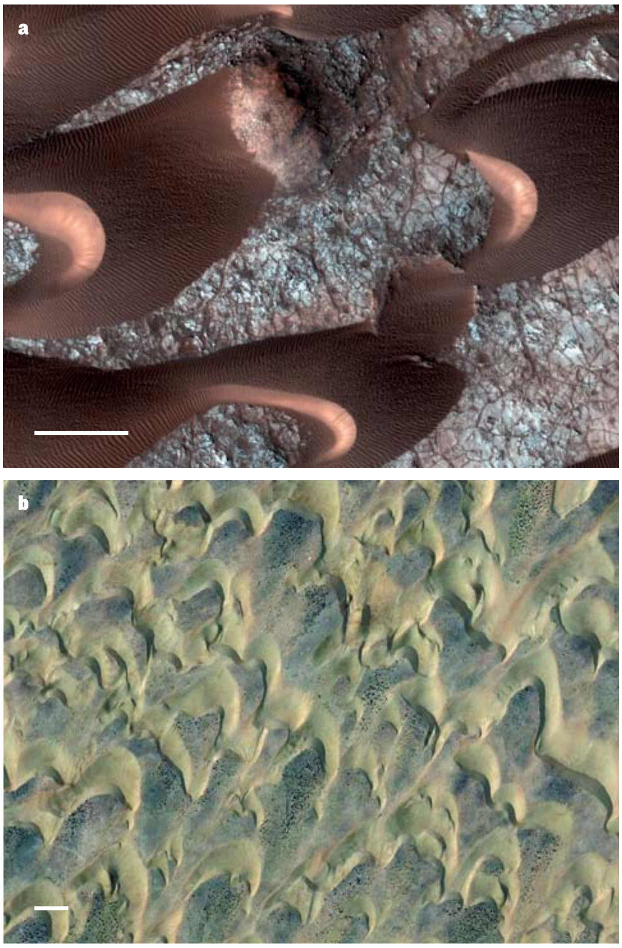


Figure 1 | Sand dunes on Mars and Earth. **a**, Bridges *et al.*[6] used high-resolution images obtained by the Martian orbital camera HiRISE to quantify movement of sand dunes and ripples over bedrock in the Martian dune field Nili Patera. **b**, The authors find that the rate at which the planet's thin atmosphere moves sand on the dunes is not much smaller than that observed for their terrestrial analogues, an example of which is shown here. Scale bars, 50 metres.

A, NASA/JPL/UNIV. ARIZONA; B, GEOEYE/GOOGLE

Bridges and colleagues arrived at this surprising conclusion by combining advances in image analysis with the astounding 25-centimetre resolution of the Martian orbital camera HiRISE (High Resolution Imaging Science Experiment) on NASA's Mars Reconnaissance Orbiter, which exceeds the resolution of any non-military satellite in orbit around Earth. By correlating high-resolution images of Nili Patera taken more than 100 days apart, the authors were able to obtain a map of sand-ripple migration across the dune field. They found that ripple displacement scaled almost perfectly with elevation on the dune, which is a telltale sign of shape-preserving dune migration. The authors then used the measured ripple dimensions to convert ripple-displacement rates to actual sand-transport rates, which correlated well with sand fluxes derived from dune displacements. This procedure resulted in an estimated average sand-transport rate of approximately seven cubic metres per spanwise metre per Earth year, which is only slightly lower than the sand flux in some dune fields on our planet (Fig. 1).

Together with the other recent studies[3–5] reporting active Martian sand transport, the surprisingly large sand fluxes estimated by Bridges *et al.* force us to view the Martian surface as a more active environment than expected. For instance, this small army of bouncing sand grains grinds away at the bedrock between the dunes much more quickly than previously thought. Moreover, the large sand flux in Nili Patera implies that this dune field could have formed in fewer than 10,000 years[6]. Because climate-altering fluctuations in Mars's orbit occur on timescales about ten times longer than that[2,7], this implies that Nili Patera, and probably other Martian dune fields, are not relicts from a previous climate with a thicker atmosphere, but could have formed in the present climate.

However, the finding that the thin Martian atmosphere transports copious amounts of sand poses more questions than it answers. In particular, atmospheric circulation models predict wind stresses that generally remain well below the threshold for sand lifting[7], even in areas where active sand transport is observed[4]. In addition, sporadic measurements by Mars landers have found that wind speeds surpass the sand-transport threshold only exceedingly rarely[1]. So the million-dollar question is: how is all this sand being moved?

Part of the answer might be that small-scale topography and convection generate strong localized winds on Mars, which cannot be accurately simulated by the coarse resolution of Martian atmospheric circulation models[8]. The transport of sand by these localized winds would also help to explain the poor correlation between areas where sand transport is observed and areas where models predict strong large-scale winds[5]. Another piece of the puzzle might be that, once a strong gust of wind starts blowing sand on Mars, the sand may well be kept adrift by moderate winds of velocities up to a factor of ten lower than that needed to initiate transport[9]. This occurs because the low Martian gravity and vertical air drag combine to make bouncing sand on Mars akin to playing golf on the Moon:

*This article and the paper[6] under discussion were published online on 9 May 2012.

particles travel much higher trajectories than on Earth, allowing them to gain substantial momentum even in light winds, so that on landing they splash up enough new particles to keep transport going at low wind speeds[9].

In addition to this question of exactly how sand is moved by the thin Martian atmosphere, Bridges *et al.* leave plenty of other exciting questions to be answered by future studies. For instance, are these high sand-transport rates typical for Mars, or does the Nili Patera dune field represent an outlier — a hotbed of aeolian activity much like the Saharan Bodélé depression on Earth[10]? And, given that many other Martian dune fields are probably inactive[1,2], which processes determine whether a dune field is active? Furthermore, on Earth, the suspended dust in dust storms is generated by the mechanical impact of blowing sand onto soils. Does the sandblasting of Martian soils similarly provide dust to Mars's many, and occasionally planet-encircling, dust storms? Or is Martian dust-lifting dominated by other processes, such as the aerodynamic lifting of low-density dust aggregates[3]? In the coming years, a widespread application of Bridges and colleagues' technique of using high-resolution satellite imagery to map Martian sand fluxes might help to provide answers to these fundamental questions. Combined with continued advances in our knowledge of the mechanics of Martian sand transport[3,9], this approach could facilitate improvements in Martian atmospheric circulation models, and drive further leaps forward in understanding our planetary neighbour. ■

Jasper Kok *is in the Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York 14853, USA.*
e-mail: jasperkok@cornell.edu

1. Zimbelman, J. R. *Geophys. Res. Lett.* **27,** 1069–1072 (2000).
2. Gardin, E. *et al. Planet. Space Sci.* **60,** 314–321 (2012).
3. Sullivan, R. *et al. J. Geophys. Res.* **113,** E06S07 (2008).
4. Chojnacki, M., Burr, D. M., Moersch, J. E. & Michaels, T. I. *J. Geophys. Res.* **116,** E00F19 (2011).
5. Bridges, N. T. *et al. Geology* **40,** 31–34 (2012).
6. Bridges, N. T. *et al. Nature* **485,** 339–342 (2012).
7. Haberle, R. M., Murphy, J. R. & Schaeffer, J. *Icarus* **161,** 66–89 (2003).
8. Fenton, L. K. & Michaels, T. I. *Mars* **5,** 159–171 (2010).
9. Kok, J. F. *Phys. Rev. Lett.* **104,** 074502 (2010).
10. Vermeesch, P. & Drake, N. *Geophys. Res. Lett.* **35,** L24404 (2008).

ATOMIC PHYSICS

Electrons get real

Strong laser fields allow electrons to tunnel out of atoms. The response of such electrons to a second laser field supports the idea that they start tunnelling at a time defined by a complex number, but exit atoms at a 'real' time. SEE LETTER P.343

MANFRED LEIN

Physicists are perfectly aware that the microscopic behaviour of electrons cannot be understood without the laws of quantum theory. Nevertheless, when scientists trace the dynamics of subatomic phenomena, they like to ask questions that are motivated by a classical, non-quantum perspective. In this spirit, Shafir *et al.*[1] report on page 343 the exact times at which electrons 'exit' atoms that are irradiated by a short flash of laser light. The existence of such an exit time is seemingly counter-intuitive, given that electrons are described by wavefunctions that extend smoothly from the inside to the outside of atoms — part of the electron is always outside the atom. In the presence of a laser field, however, there is a continuous outward flow of electron density, which Shafir and colleagues have decomposed into different electron trajectories, assigning each trajectory an experimentally determined starting time.

The emission of electrons from atoms in Shafir and colleagues' experiments is a consequence of quantum tunnelling. The applied laser field changes the potential-energy profile experienced by the electron, forming a finite barrier that is impenetrable to classical Newtonian particles, but which can be tunnelled across by electrons. A similar process forms the basis of scanning tunnelling microscopy: electrons tunnel between the surface of the object under study and the tip of the microscope. Tunnelling occurs because electron wavefunctions encompass both sides of a potential barrier (Fig. 1a); so what is the meaning of an exit time?

Before answering that question, one must realize that the authors did not detect electrons in their experiments. Instead, they recorded the light released on the return of an emitted electron to its parent ion. Light release occurs because the electric field in a laser pulse reverses direction periodically. This means that, about 1 femtosecond (10^{-15} seconds) after an electron has tunnelled out of an atom, the laser's force pushes it back towards the resulting ion (Fig. 1b). If the electron and ion recombine to form the same bound state that existed before ionization, then a photon is emitted[2]. Because the photon's frequency (and therefore its energy) is much higher than that of the incident laser light, the photon-forming process is called high-harmonic generation.

To resolve the electron emission in time, Shafir *et al.* perturbed electrons emitted from helium atoms with a second, weak probe field acting perpendicular to the main laser field. To picture the experiment, imagine a game

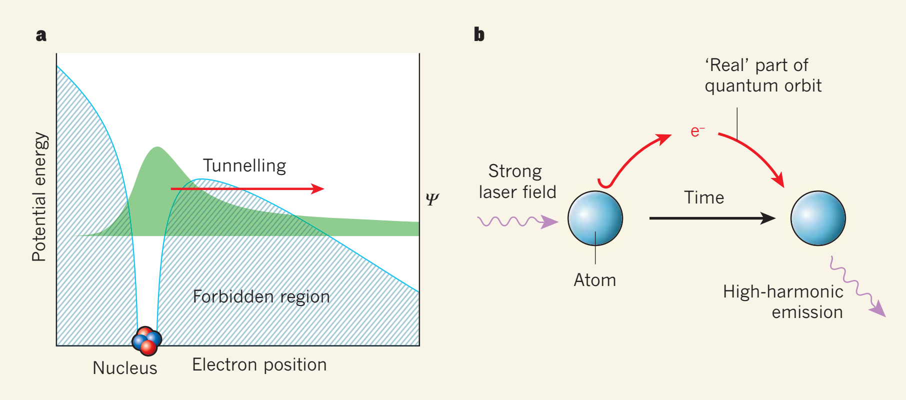


Figure 1 | Quantum tunnelling and high-harmonic generation. a, The blue line depicts the potential-energy profile that binds an electron in a laser-irradiated atom. The atom's nucleus is at the profile's minimum. If the electron were a classical Newtonian particle, it could not enter the shaded 'forbidden' regions and would be trapped in the atom. But electrons are quantum-mechanical objects whose probability distributions in space are described by wavefunctions (Ψ, such as the one shown in green). Because wavefunctions extend through the right-hand forbidden area, electrons may tunnel out of the atom. **b,** In the phenomenon of high-harmonic generation, a laser field accelerates an electron (e^-) that has tunnelled out of an atom away from the resulting ion, then directs it back again. Recombination of the electron with the parent ion generates a high-energy photon (a high-harmonic emission). Shafir and colleagues' report[1] suggests that high-harmonic emissions from helium atoms are described by 'quantum orbits'. This means that tunnelling proceeds in imaginary time (the imaginary part of time as defined by a complex number), but the electron moves as a classical particle in 'real' time once it has exited the atom. At the start of its real-time journey, the electron counter-intuitively moves towards the parent ion.

particles travel much higher trajectories than on Earth, allowing them to gain substantial momentum even in light winds, so that on landing they splash up enough new particles to keep transport going at low wind speeds[9].

In addition to this question of exactly how sand is moved by the thin Martian atmosphere, Bridges *et al.* leave plenty of other exciting questions to be answered by future studies. For instance, are these high sand-transport rates typical for Mars, or does the Nili Patera dune field represent an outlier — a hotbed of aeolian activity much like the Saharan Bodélé depression on Earth[10]? And, given that many other Martian dune fields are probably inactive[1,2], which processes determine whether a dune field is active? Furthermore, on Earth, the suspended dust in dust storms is generated by the mechanical impact of blowing sand onto soils. Does the sandblasting of Martian soils similarly provide dust to Mars's many, and occasionally planet-encircling, dust storms? Or is Martian dust-lifting dominated by other processes, such as the aerodynamic lifting of low-density dust aggregates[3]? In the coming years, a widespread application of Bridges and colleagues' technique of using high-resolution satellite imagery to map Martian sand fluxes might help to provide answers to these fundamental questions. Combined with continued advances in our knowledge of the mechanics of Martian sand transport[3,9], this approach could facilitate improvements in Martian atmospheric circulation models, and drive further leaps forward in understanding our planetary neighbour. ■

Jasper Kok *is in the Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York 14853, USA.*
e-mail: jasperkok@cornell.edu

1. Zimbelman, J. R. *Geophys. Res. Lett.* **27,** 1069–1072 (2000).
2. Gardin, E. *et al. Planet. Space Sci.* **60,** 314–321 (2012).
3. Sullivan, R. *et al. J. Geophys. Res.* **113,** E06S07 (2008).
4. Chojnacki, M., Burr, D. M., Moersch, J. E. & Michaels, T. I. *J. Geophys. Res.* **116,** E00F19 (2011).
5. Bridges, N. T. *et al. Geology* **40,** 31–34 (2012).
6. Bridges, N. T. *et al. Nature* **485,** 339–342 (2012).
7. Haberle, R. M., Murphy, J. R. & Schaeffer, J. *Icarus* **161,** 66–89 (2003).
8. Fenton, L. K. & Michaels, T. I. *Mars* **5,** 159–171 (2010).
9. Kok, J. F. *Phys. Rev. Lett.* **104,** 074502 (2010).
10. Vermeesch, P. & Drake, N. *Geophys. Res. Lett.* **35,** L24404 (2008).

ATOMIC PHYSICS

Electrons get real

Strong laser fields allow electrons to tunnel out of atoms. The response of such electrons to a second laser field supports the idea that they start tunnelling at a time defined by a complex number, but exit atoms at a 'real' time. **SEE LETTER P.343**

MANFRED LEIN

Physicists are perfectly aware that the microscopic behaviour of electrons cannot be understood without the laws of quantum theory. Nevertheless, when scientists trace the dynamics of subatomic phenomena, they like to ask questions that are motivated by a classical, non-quantum perspective. In this spirit, Shafir *et al.*[1] report on page 343 the exact times at which electrons 'exit' atoms that are irradiated by a short flash of laser light. The existence of such an exit time is seemingly counter-intuitive, given that electrons are described by wavefunctions that extend smoothly from the inside to the outside of atoms — part of the electron is always outside the atom. In the presence of a laser field, however, there is a continuous outward flow of electron density, which Shafir and colleagues have decomposed into different electron trajectories, assigning each trajectory an experimentally determined starting time.

The emission of electrons from atoms in Shafir and colleagues' experiments is a consequence of quantum tunnelling. The applied laser field changes the potential-energy profile experienced by the electron, forming a finite barrier that is impenetrable to classical Newtonian particles, but which can be tunnelled across by electrons. A similar process forms the basis of scanning tunnelling microscopy: electrons tunnel between the surface of the object under study and the tip of the microscope. Tunnelling occurs because electron wavefunctions encompass both sides of a potential barrier (Fig. 1a); so what is the meaning of an exit time?

Before answering that question, one must realize that the authors did not detect electrons in their experiments. Instead, they recorded the light released on the return of an emitted electron to its parent ion. Light release occurs because the electric field in a laser pulse reverses direction periodically. This means that, about 1 femtosecond (10^{-15} seconds) after an electron has tunnelled out of an atom, the laser's force pushes it back towards the resulting ion (Fig. 1b). If the electron and ion recombine to form the same bound state that existed before ionization, then a photon is emitted[2]. Because the photon's frequency (and therefore its energy) is much higher than that of the incident laser light, the photon-forming process is called high-harmonic generation.

To resolve the electron emission in time, Shafir *et al.* perturbed electrons emitted from helium atoms with a second, weak probe field acting perpendicular to the main laser field. To picture the experiment, imagine a game

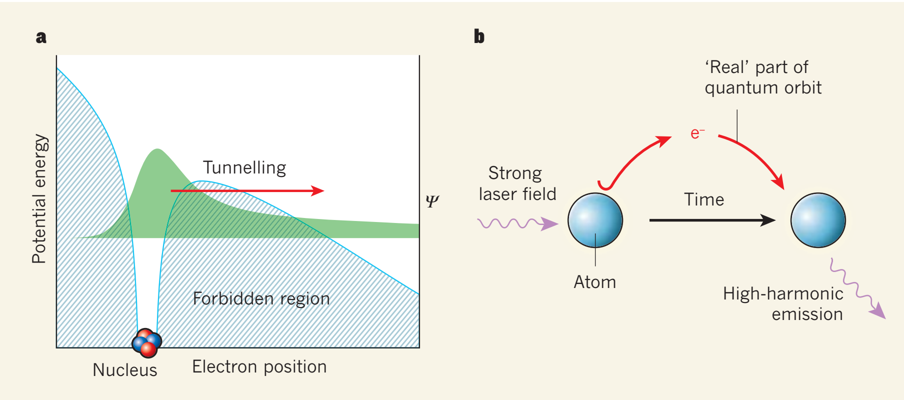


Figure 1 | Quantum tunnelling and high-harmonic generation. **a**, The blue line depicts the potential-energy profile that binds an electron in a laser-irradiated atom. The atom's nucleus is at the profile's minimum. If the electron were a classical Newtonian particle, it could not enter the shaded 'forbidden' regions and would be trapped in the atom. But electrons are quantum-mechanical objects whose probability distributions in space are described by wavefunctions (Ψ, such as the one shown in green). Because wavefunctions extend through the right-hand forbidden area, electrons may tunnel out of the atom. **b**, In the phenomenon of high-harmonic generation, a laser field accelerates an electron (e^-) that has tunnelled out of an atom away from the resulting ion, then directs it back again. Recombination of the electron with the parent ion generates a high-energy photon (a high-harmonic emission). Shafir and colleagues' report[1] suggests that high-harmonic emissions from helium atoms are described by 'quantum orbits'. This means that tunnelling proceeds in imaginary time (the imaginary part of time as defined by a complex number), but the electron moves as a classical particle in 'real' time once it has exited the atom. At the start of its real-time journey, the electron counter-intuitively moves towards the parent ion.

in which people (the atoms) repeatedly and regularly throw balls (the electrons) vertically into the air and then catch them. Even without seeing the players, one can discern a successful capture by hearing them cheer "Yeah" (photon emission).

Now suppose you could blow a crosswind (the probe field) over the heads of the players. If the crosswind blows when the ball is in the air, it pushes the ball sideways. In this case, the ball lands some distance away from its starting point so that the player cannot catch it, and there is no cheer. By applying the crosswind at various times and then listening for the presence or absence of cheers, one can determine the precise throwing times (the ionization times).

In their experiments, Shafir *et al.* used a probe field that oscillated at twice the frequency of the main field, and monitored the intensity of the emitted harmonics as the researchers varied the temporal shift between the two fields (the time elapsing between a maximum of the main field and a maximum of the probe field). But for a robust analysis, they also needed to measure electron departure and return times independently, which meant that they had to take things further. In our analogous game, a crosswind whose direction alternates could affect the ball so much that it unexpectedly hits the player from the side, making him shout "Yikes" instead of "Yeah". Similarly, Shafir *et al.* deduced the angle of electron return in their experiments by detecting specific photon emissions known as even-order harmonics. These emissions, whose frequencies are even multiples of the main laser's frequency, were generated when released electrons hit their parent ions 'from the side' — at an angle to the main field.

Of particular note, the authors found that every harmonic emission frequency has its own ionization time, all of which fall within a range of 200 attoseconds (1 attosecond is 10^{-18} s). The superposition of the many different associated electron trajectories forms a quantum-mechanical wave packet — a short 'pulse' of travelling wave activity — for emitted electrons. The observed ionization times[1] are conceptually different from the extremely small tunnelling delay time (tens of attoseconds at most) reported in a previous study of helium atoms[3], which measured the delay between the maximum of the applied oscillating laser field and the most likely time of electron departure.

One striking result is the excellent agreement of Shafir and colleagues' findings with a model of high-harmonic generation that is well known to atomic physicists — the quantum orbit model[4]. Put simply, the model states that an electron trajectory begins with negative kinetic energy at an instant of time defined by a complex number. Just the imaginary part of time changes for the electron as it tunnels through a potential barrier; time becomes real-valued only at the exit of the tunnel. This real time is the exit time measured by Shafir and colleagues. It is the time at which the electron starts to feel the effect of the probe field.

In molecules, high-harmonic generation often involves contributions from different 'channels' — that is, not only from the most weakly bound electrons of atoms, but also from more tightly bound ones[5,6]. Shafir *et al.*[1] report that small differences in ionization times from two channels in carbon dioxide are, in principle, measurable. The characteristic differences are observable when the channels interfere nearly destructively. However, this scenario corresponds to a small range of the emitted spectrum, and generates a low number of photons. Determining reliable, channel-dependent ionization times will therefore be extremely challenging.

One limitation of Shafir and colleagues' study is that they measure only how ionization times and return times vary with harmonic frequency. But the absolute timing of ionization is of substantial interest too, because it is related to the tunnelling delay time and it may influence the absolute timing of the harmonic emission. It remains to be seen how well the authors' technique will work for mixtures of gases, in which differences between atomic species may complicate harmonic emission profiles. Extending the present study to such a case may reveal interference between emissions from different gases in the same way that different channels in the same molecule interfere with each other.

Accurate knowledge of the electron excursion times (the difference between return and ionization times) during high-harmonic generation is vital for our understanding of the many ultrafast experiments[5,7] in which ionization triggers a dynamic process, and in which recombination of the resulting ion with the electron takes a snapshot of that process. Unlike the classical approach[8] to ultrafast time-resolved chemistry, in which reactions are initiated using a 'pump' light pulse and a separate probe pulse is used to monitor reaction evolution, high-harmonic generation combines the pump and probe steps into just one shot. By facilitating the real-time observation of attosecond electron dynamics, this approach will increasingly compete with ultrafast spectroscopic methods in which molecules are directly probed by attosecond light pulses[9]. ■

Manfred Lein *is at the Centre for Quantum Engineering and Space-Time Technology and at the Institute for Theoretical Physics, Leibniz Universität Hannover, 30167 Hanover, Germany.*
e-mail: lein@itp.uni-hannover.de

1. Shafir, D. *et al. Nature* **485,** 343–346 (2012).
2. Corkum, P. B. *Phys. Rev. Lett.* **71,** 1994–1997 (1993).
3. Eckle, P. *et al. Science* **322,** 1525–1529 (2008).
4. Salières, P. *et al. Science* **292,** 902–905 (2001).
5. Smirnova, O. *et al. Nature* **460,** 972–977 (2009).
6. McFarland, B. K., Farrell, J. P., Bucksbaum, P. H. & Gühr, M. *Science* **322,** 1232–1235 (2008).
7. Baker, S. *et al. Science* **312,** 424–427 (2006).
8. Zewail, A. H. *Science* **242,** 1645–1653 (1988).
9. Krausz, F. & Ivanov, M. *Rev. Mod. Phys.* **81,** 163–234 (2009).

STRUCTURAL BIOLOGY

How opioid drugs bind to receptors

The search for safe, non-addictive versions of morphine and other opioid drugs has just received a boost with the solving of the crystal structures of the receptors to which the drugs bind. SEE ARTICLES P.321 & P.327, LETTERS P.395 & P.400

MARTA FILIZOLA & LAKSHMI A. DEVI

Opioid drugs such as morphine and codeine are powerful painkillers, but an assortment of adverse side effects limits their effective medical use. These drugs can also produce pronounced euphoria, which has led to the recreational use of common prescription painkillers. Addiction to prescription opioids is currently one of the most severe forms of drug abuse[1], a fact that raises significant public-health concerns and highlights a pressing need for the development of safer painkillers. In this issue, four papers[2–5] report crystal structures that provide the first direct evidence for the binding mode of opioids to their receptors. This information will be invaluable for research aimed at finding opioid drugs that lack the adverse side effects.

Opioid receptors (ORs) are members of the superfamily of G-protein-coupled receptors (GPCRs). The traditional model of OR signalling proposes that the binding of a ligand molecule (an opioid) to a receptor activates an associated G protein, which, in turn, triggers a biological response. Widely distributed in the brain and in the peripheral nervous system, the four types of OR are μ-OR, δ-OR, κ-OR and the nociceptin/orphanin FQ peptide receptor. These receptors represent prominent

in which people (the atoms) repeatedly and regularly throw balls (the electrons) vertically into the air and then catch them. Even without seeing the players, one can discern a successful capture by hearing them cheer "Yeah" (photon emission).

Now suppose you could blow a crosswind (the probe field) over the heads of the players. If the crosswind blows when the ball is in the air, it pushes the ball sideways. In this case, the ball lands some distance away from its starting point so that the player cannot catch it, and there is no cheer. By applying the crosswind at various times and then listening for the presence or absence of cheers, one can determine the precise throwing times (the ionization times).

In their experiments, Shafir *et al.* used a probe field that oscillated at twice the frequency of the main field, and monitored the intensity of the emitted harmonics as the researchers varied the temporal shift between the two fields (the time elapsing between a maximum of the main field and a maximum of the probe field). But for a robust analysis, they also needed to measure electron departure and return times independently, which meant that they had to take things further. In our analogous game, a crosswind whose direction alternates could affect the ball so much that it unexpectedly hits the player from the side, making him shout "Yikes" instead of "Yeah". Similarly, Shafir *et al.* deduced the angle of electron return in their experiments by detecting specific photon emissions known as even-order harmonics. These emissions, whose frequencies are even multiples of the main laser's frequency, were generated when released electrons hit their parent ions 'from the side' — at an angle to the main field.

Of particular note, the authors found that every harmonic emission frequency has its own ionization time, all of which fall within a range of 200 attoseconds (1 attosecond is 10^{-18} s). The superposition of the many different associated electron trajectories forms a quantum-mechanical wave packet — a short 'pulse' of travelling wave activity — for emitted electrons. The observed ionization times[1] are conceptually different from the extremely small tunnelling delay time (tens of attoseconds at most) reported in a previous study of helium atoms[3], which measured the delay between the maximum of the applied oscillating laser field and the most likely time of electron departure.

One striking result is the excellent agreement of Shafir and colleagues' findings with a model of high-harmonic generation that is well known to atomic physicists — the quantum orbit model[4]. Put simply, the model states that an electron trajectory begins with negative kinetic energy at an instant of time defined by a complex number. Just the imaginary part of time changes for the electron as it tunnels through a potential barrier; time becomes real-valued only at the exit of the tunnel. This real time is the exit time measured by Shafir and colleagues. It is the time at which the electron starts to feel the effect of the probe field.

In molecules, high-harmonic generation often involves contributions from different 'channels' — that is, not only from the most weakly bound electrons of atoms, but also from more tightly bound ones[5,6]. Shafir *et al.*[1] report that small differences in ionization times from two channels in carbon dioxide are, in principle, measurable. The characteristic differences are observable when the channels interfere nearly destructively. However, this scenario corresponds to a small range of the emitted spectrum, and generates a low number of photons. Determining reliable, channel-dependent ionization times will therefore be extremely challenging.

One limitation of Shafir and colleagues' study is that they measure only how ionization times and return times vary with harmonic frequency. But the absolute timing of ionization is of substantial interest too, because it is related to the tunnelling delay time and it may influence the absolute timing of the harmonic emission. It remains to be seen how well the authors' technique will work for mixtures of gases, in which differences between atomic species may complicate harmonic emission profiles. Extending the present study to such a case may reveal interference between emissions from different gases in the same way that different channels in the same molecule interfere with each other.

Accurate knowledge of the electron excursion times (the difference between return and ionization times) during high-harmonic generation is vital for our understanding of the many ultrafast experiments[5,7] in which ionization triggers a dynamic process, and in which recombination of the resulting ion with the electron takes a snapshot of that process. Unlike the classical approach[8] to ultrafast time-resolved chemistry, in which reactions are initiated using a 'pump' light pulse and a separate probe pulse is used to monitor reaction evolution, high-harmonic generation combines the pump and probe steps into just one shot. By facilitating the real-time observation of attosecond electron dynamics, this approach will increasingly compete with ultrafast spectroscopic methods in which molecules are directly probed by attosecond light pulses[9]. ■

Manfred Lein *is at the Centre for Quantum Engineering and Space-Time Technology and at the Institute for Theoretical Physics, Leibniz Universität Hannover, 30167 Hanover, Germany.*
e-mail: lein@itp.uni-hannover.de

1. Shafir, D. *et al. Nature* **485,** 343–346 (2012).
2. Corkum, P. B. *Phys. Rev. Lett.* **71,** 1994–1997 (1993).
3. Eckle, P. *et al. Science* **322,** 1525–1529 (2008).
4. Salières, P. *et al. Science* **292,** 902–905 (2001).
5. Smirnova, O. *et al. Nature* **460,** 972–977 (2009).
6. McFarland, B. K., Farrell, J. P., Bucksbaum, P. H. & Gühr, M. *Science* **322,** 1232–1235 (2008).
7. Baker, S. *et al. Science* **312,** 424–427 (2006).
8. Zewail, A. H. *Science* **242,** 1645–1653 (1988).
9. Krausz, F. & Ivanov, M. *Rev. Mod. Phys.* **81,** 163–234 (2009).

STRUCTURAL BIOLOGY

How opioid drugs bind to receptors

The search for safe, non-addictive versions of morphine and other opioid drugs has just received a boost with the solving of the crystal structures of the receptors to which the drugs bind. SEE ARTICLES P.321 & P.327, LETTERS P.395 & P.400

MARTA FILIZOLA & LAKSHMI A. DEVI

Opioid drugs such as morphine and codeine are powerful painkillers, but an assortment of adverse side effects limits their effective medical use. These drugs can also produce pronounced euphoria, which has led to the recreational use of common prescription painkillers. Addiction to prescription opioids is currently one of the most severe forms of drug abuse[1], a fact that raises significant public-health concerns and highlights a pressing need for the development of safer painkillers. In this issue, four papers[2–5] report crystal structures that provide the first direct evidence for the binding mode of opioids to their receptors. This information will be invaluable for research aimed at finding opioid drugs that lack the adverse side effects.

Opioid receptors (ORs) are members of the superfamily of G-protein-coupled receptors (GPCRs). The traditional model of OR signalling proposes that the binding of a ligand molecule (an opioid) to a receptor activates an associated G protein, which, in turn, triggers a biological response. Widely distributed in the brain and in the peripheral nervous system, the four types of OR are μ-OR, δ-OR, κ-OR and the nociceptin/orphanin FQ peptide receptor. These receptors represent prominent

targets not only for painkillers, but also for antidepressants, anti-addiction medications and anti-anxiety drugs.

The papers in this issue[2–5] present the long-awaited, high-resolution crystal structures of all four ORs in ligand-bound conformations. The ligands are all antagonists (receptor blockers), which means that the structures depict inactive states of the receptors. These crystal structures are the latest to have been obtained using revolutionary technologies — including the replacement of part of the receptors with another protein, such as T4 lysozyme[6,7], to facilitate receptor crystallization — that have enabled successful structural determination of several GPCRs. Such proteins were once intractable to crystallography.

The four OR structures reveal several evolutionarily conserved ligand–receptor interactions in the receptors' binding pockets, which are contained within the seven transmembrane helices (designated TM1–7) of the receptors. For instance, several amino-acid residues at the same positions in TM3, TM6 and TM7 form interactions with the chemical moieties of ligands that are responsible for opioid efficacy — the 'message' region of the ligands. By contrast, the chemical moieties responsible for opioid selectivity — the 'address' region — occupy one of two different areas of the binding pocket, depending on the type of opioid. Specifically, the addresses of classical opioids, which contain the 'morphinan' chemical structure, interact with TM6 and/or TM7, whereas the corresponding regions of the other opioids studied are positioned between TM2 and TM3 of the receptor (Fig. 1), forming interactions mostly with those helices, but also with TM7. Accordingly, Wu and colleagues suggest[3] that the message–address hypothesis of opioid binding may not apply uniformly to all opioid ligands.

The transmembrane structures of the four ORs are very similar to each other, as expected given that the amino-acid sequences of these structures are also very similar (homologous, to use the jargon). More surprisingly, the structures of non-homologous loop regions, such as the long, extracellular loop region ECL2, are also very alike. Notably, the ECL2 structure of the ORs is similar to that[8] of CXCR4 — another GPCR that, like the ORs, binds both peptides and small molecules. This shared, 'β-hairpin' loop structure creates a wide opening that allows ligands unobstructed access to the primary binding pocket within the transmembrane region. Manglik *et al.* suggest[4] that this might explain why the effects of most opioid drugs are highly potent yet rapidly reversible.

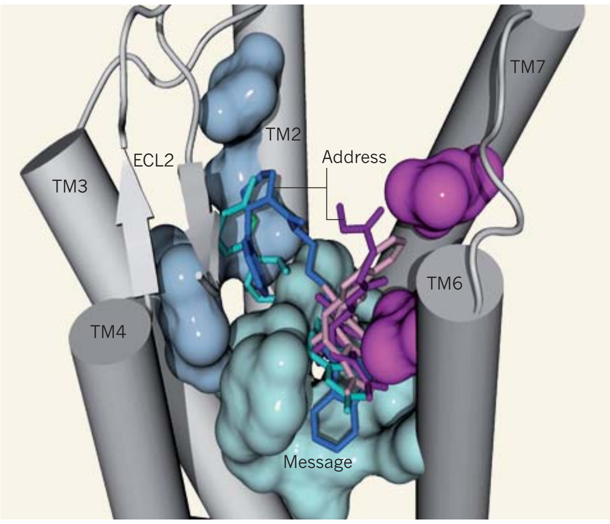


Figure 1 | Binding mode of opioids at their receptors. The structures of the four types of opioid receptor, each in complex with a different opioid antagonist, have been solved[2–5]. A side view of one of the structures — that of the nociceptin/orphanin FQ peptide (NOP) receptor — is depicted to show features shared by all four receptor types. Only five of the seven transmembrane helices (TM1–7) are shown (grey cylinders). ECL2 is a β-hairpin loop region; the arrows represent β-sheets. The four antagonists used in the studies are depicted as stick representations in the NOP receptor's binding pocket. The cyan surface indicates the amino-acid residues from TM3, TM6 and TM7 that interact with the ligands' 'message' regions, responsible for a ligand's efficacy. The magenta surfaces indicate the residues from TM6 and/or TM7 that interact with the 'address' region — responsible for opioid selectivity — of classical ligands, which contain the 'morphinan' chemical structure. The light-blue surfaces represent residues from TM2 and TM3 that interact with the address region of non-classical opioids.

Analysis of the OR crystal structures also reveals an unexpected outward displacement of the extracellular half of TM1 away from the long axis of κ-OR (ref. 3), compared with the other opioid receptors[2,4,5] and CXCR4 (ref. 8). However, as previously noted[9,10] in the case of another GPCR — the β1-adrenergic receptor — different conformations of TM1 (and TM6) can be identified in inactive structures as a result of different crystal-packing interactions and/or crystallization conditions. In other words, the unusual conformation of TM1 in κ-OR may simply be one of many conformations that could have been adopted by the helix. This is an important point, as it reflects the intrinsic dynamic nature of GPCRs. Moreover, it reminds us that crystal structures of GPCRs are single, static snapshots of receptors stripped of their natural lipid environment, and might therefore offer limited mechanistic insight.

Evidence suggests[11] that the most addictive opioids promote OR interactions with their G proteins more strongly than with arrestin, another cellular signalling protein. To develop drugs that retain the therapeutic action of opioids but not the unwanted side effects, it is therefore crucial to understand the specific receptor conformations that opioids stabilize to selectively activate signalling pathways. This important aspect of ligand binding to ORs is not captured by the recent crystal structures, and should be the subject of future research.

There is also compelling evidence[12,13] that different types of OR associate with each other, or with other GPCR subtypes, to form dimers and oligomers, and that this changes the signalling properties of the ORs, thereby adding an additional level of complexity to an already multifaceted problem. Manglik and colleagues' structure[4] of the μ-OR shows tightly associated pairs of receptor molecules, held together predominantly by highly complementary interactions involving TM5 and TM6. The researchers speculate that this pairing might regulate the signalling of the receptor. A similar interaction was noted[8] in the structure of CXCR4, but is not found in the other OR structures[2,3,5].

By contrast, the κ-OR structure shows a dimeric arrangement involving interactions of TM1, TM2 and helix 8 (H8), which is similar to the alternative, less compact crystal packing seen in the μ-OR structure. The proposed roles of the TM5–TM6 and TM1–TM2–H8 interfaces are only two of several working hypotheses of functionally relevant receptor–receptor interactions that need to be addressed to enable investigators to examine the role of dimerization (or oligomerization) in the signalling of ORs. The quest for functionally relevant oligomerization interfaces therefore continues.

These crystal structures[2–5] of inactive ORs will contribute crucial information to a broad range of therapeutic areas, including those focused on pain, addiction and mental disorders. Future crystal structures of active ORs in complex with different signalling proteins could provide necessary — although not sufficient — information for elucidating the mechanisms underlying receptor function. A complete understanding will also require the integration of experimental and computational strategies that allow the study of receptors in a natural lipid environment — necessary to obtain rigorous mechanistic insight, at the molecular level, into the ligand-induced conformation selection, spatio-temporal organization and dynamics of OR complexes. The challenge will then be to translate that knowledge from bench to bedside, by fine-tuning OR signalling towards therapeutic pathways, and away from those that mediate adverse side effects. ■

Marta Filizola *is in the Department of Structural and Chemical Biology, and* **Lakshmi A. Devi** *is in the Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, New York 10029, USA.*
e-mails: marta.filizola@mssm.edu; lakshmi.devi@mssm.edu

1. *Results from the 2009 National Survey on Drug Use and Health: Volume I. Summary of National Findings* (Office of Applied Studies, Rockville, MD, 2010).
2. Thompson, A. A. *et al. Nature* **485,** 395–399 (2012).
3. Wu, H. *et al. Nature* **485,** 327–332 (2012).
4. Manglik, A. *et al. Nature* **485,** 321–326 (2012).
5. Granier, S. *et al. Nature* **485,** 400–404 (2012).
6. Rosenbaum, D. M. *et al. Science* **318,** 1266–1273 (2007).
7. Cherezov, V., Abola, E. & Stevens, R. C. *Meth. Mol. Biol.* **654,** 141–168 (2010).
8. Wu, B. *et al. Science* **330,** 1066–1071 (2010).
9. Warne, T. *et al. Nature* **454,** 486–491 (2008).
10. Moukhametzianov, R. *et al. Proc. Natl Acad. Sci. USA* **108,** 8228–8232 (2011).
11. Molinari, P. *et al. J. Biol. Chem.* **285,** 12522–12535 (2010).
12. Rozenfeld, R. & Devi, L. A. *Trends Pharmacol Sci.* **31,** 124–130 (2010).
13. van Rijn, R. M., Whistler, J. L. & Waldhoer, M. *Curr. Opin. Pharmacol.* **10,** 73–79 (2010).

NEUROSCIENCE

Brain-controlled robot grabs attention

Restoring voluntary actions to paralysed patients is an ambition of neural-interface research. A study shows that people with tetraplegia can use brain control of a robotic arm to reach and grasp objects. SEE LETTER P.372

ANDREW JACKSON

Most of us take for granted our effortless ability to interact with objects. When we are thirsty and reach for a cup, electrical signals stream from the brain through the spinal cord, instructing our muscles to move. However, a disruption of the nerve pathways along which these signals travel can cause paralysis, with devastating consequences for the person's quality of life. So there is growing interest in technologies that allow the brain to bypass nerve injuries and communicate directly with the environment. Neural-interface systems, also known as brain–machine interfaces, detect electrical signals in the brain and use them to control external assistive devices. The first results from a clinical trial of 'BrainGate', a neural interface that enabled a patient paralysed by a spinal-cord injury to move a computer cursor, were published[1] in 2006. On page 372 of this issue, Hochberg *et al.*[2] now report that two people with long-standing paralysis can control the reaching and grasping actions of a robotic arm using BrainGate. One of the participants was even able to drink from a bottle using the robotic arm, something she had not been able to do with her own limb since a stroke nearly 15 years ago.

To access brain signals, BrainGate uses thin silicon electrodes surgically inserted a few millimetres into the primary motor cortex, a part of the brain that controls movements. Remarkably, neurons in this area responded when the patients imagined controlling the robotic arm, although both of them had lost the use of their limbs many years earlier. During a calibration phase, the researchers constructed a 'decoder' that translated participants' intentions into three-dimensional movements and into a closing of the robotic hand. They then tested the participants' ability to reach for and grasp foam balls presented in front of them.

Although the speed and accuracy of the robot's movements fell well short of those of natural arm control, the participants successfully touched the foam balls on 49% to 95% of attempts across multiple sessions with two different robot designs. What's more, about two-thirds of successful reaches resulted in correct grasping. The authors further established the efficacy of brain control by one participant in a bottle-grasping and drinking task, demonstrating that a neural-interface system can perform actions that are useful in daily life.

Apart from being one of only a handful of studies to use indwelling electrodes for neural interfacing in humans, Hochberg and colleagues' work is notable in that one patient had had the implanted electrodes for more than five years. Although several techniques (such as electroencephalography) can record signals from the brain in a non-invasive manner, it is generally thought that electrodes positioned inside the brain convey more information. However, as well as the risks associated with surgery, a disadvantage of such implants is the potential for scar tissue to form around the electrodes, which can result

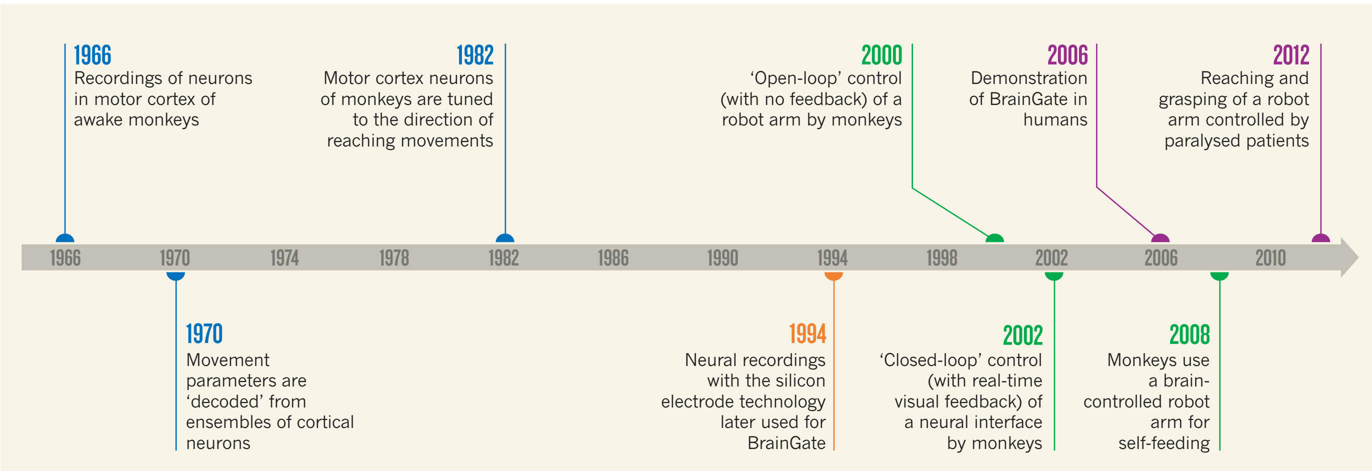


Figure 1 | Within reach. Hochberg *et al.*[2] show that people with tetraplegia can use a neural device, known as BrainGate, to control a robotic arm for reaching and grasping objects. This work builds on decades of previous research on the neural mechanisms that control arm movements[13–15] (blue), on electrode development[16] (orange) and on neural interfaces in monkeys[3–6] (green), which opened the way to studies in humans[1,2] (purple).

Marta Filizola *is in the Department of Structural and Chemical Biology, and* **Lakshmi A. Devi** *is in the Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, New York 10029, USA.*
e-mails: marta.filizola@mssm.edu; lakshmi.devi@mssm.edu

1. *Results from the 2009 National Survey on Drug Use and Health: Volume I. Summary of National Findings* (Office of Applied Studies, Rockville, MD, 2010).
2. Thompson, A. A. *et al. Nature* **485,** 395–399 (2012).
3. Wu, H. *et al. Nature* **485,** 327–332 (2012).
4. Manglik, A. *et al. Nature* **485,** 321–326 (2012).
5. Granier, S. *et al. Nature* **485,** 400–404 (2012).
6. Rosenbaum, D. M. *et al. Science* **318,** 1266–1273 (2007).
7. Cherezov, V., Abola, E. & Stevens, R. C. *Meth. Mol. Biol.* **654,** 141–168 (2010).
8. Wu, B. *et al. Science* **330,** 1066–1071 (2010).
9. Warne, T. *et al. Nature* **454,** 486–491 (2008).
10. Moukhametzianov, R. *et al. Proc. Natl Acad. Sci. USA* **108,** 8228–8232 (2011).
11. Molinari, P. *et al. J. Biol. Chem.* **285,** 12522–12535 (2010).
12. Rozenfeld, R. & Devi, L. A. *Trends Pharmacol Sci.* **31,** 124–130 (2010).
13. van Rijn, R. M., Whistler, J. L. & Waldhoer, M. *Curr. Opin. Pharmacol.* **10,** 73–79 (2010).

NEUROSCIENCE

Brain-controlled robot grabs attention

Restoring voluntary actions to paralysed patients is an ambition of neural-interface research. A study shows that people with tetraplegia can use brain control of a robotic arm to reach and grasp objects. **See Letter p.372**

ANDREW JACKSON

Most of us take for granted our effortless ability to interact with objects. When we are thirsty and reach for a cup, electrical signals stream from the brain through the spinal cord, instructing our muscles to move. However, a disruption of the nerve pathways along which these signals travel can cause paralysis, with devastating consequences for the person's quality of life. So there is growing interest in technologies that allow the brain to bypass nerve injuries and communicate directly with the environment. Neural-interface systems, also known as brain–machine interfaces, detect electrical signals in the brain and use them to control external assistive devices. The first results from a clinical trial of 'BrainGate', a neural interface that enabled a patient paralysed by a spinal-cord injury to move a computer cursor, were published[1] in 2006. On page 372 of this issue, Hochberg *et al.*[2] now report that two people with long-standing paralysis can control the reaching and grasping actions of a robotic arm using BrainGate. One of the participants was even able to drink from a bottle using the robotic arm, something she had not been able to do with her own limb since a stroke nearly 15 years ago.

To access brain signals, BrainGate uses thin silicon electrodes surgically inserted a few millimetres into the primary motor cortex, a part of the brain that controls movements. Remarkably, neurons in this area responded when the patients imagined controlling the robotic arm, although both of them had lost the use of their limbs many years earlier. During a calibration phase, the researchers constructed a 'decoder' that translated participants' intentions into three-dimensional movements and into a closing of the robotic hand. They then tested the participants' ability to reach for and grasp foam balls presented in front of them.

Although the speed and accuracy of the robot's movements fell well short of those of natural arm control, the participants successfully touched the foam balls on 49% to 95% of attempts across multiple sessions with two different robot designs. What's more, about two-thirds of successful reaches resulted in correct grasping. The authors further established the efficacy of brain control by one participant in a bottle-grasping and drinking task, demonstrating that a neural-interface system can perform actions that are useful in daily life.

Apart from being one of only a handful of studies to use indwelling electrodes for neural interfacing in humans, Hochberg and colleagues' work is notable in that one patient had had the implanted electrodes for more than five years. Although several techniques (such as electroencephalography) can record signals from the brain in a non-invasive manner, it is generally thought that electrodes positioned inside the brain convey more information. However, as well as the risks associated with surgery, a disadvantage of such implants is the potential for scar tissue to form around the electrodes, which can result

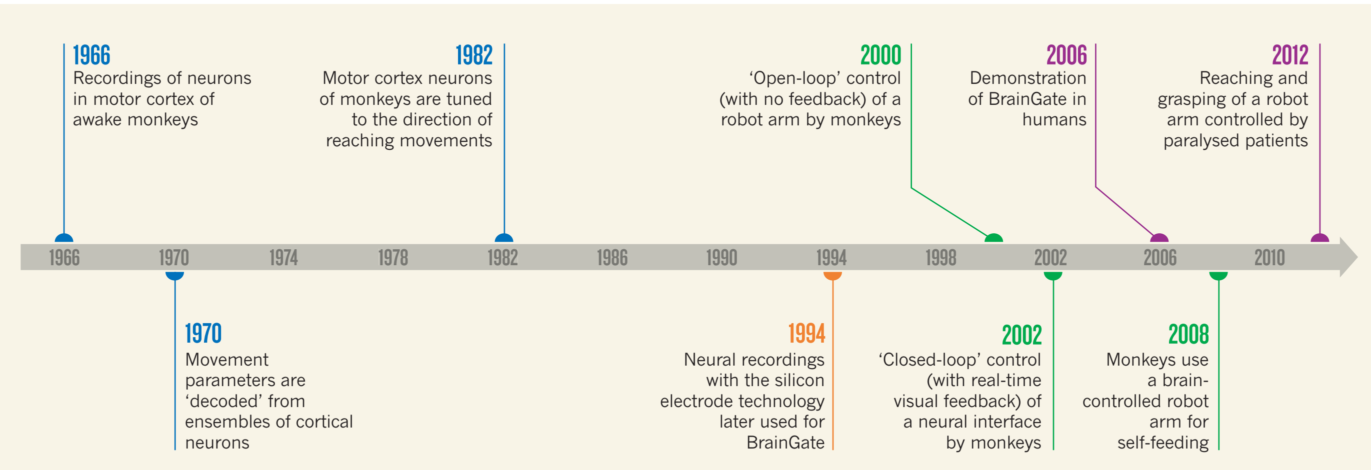


Figure 1 | Within reach. Hochberg *et al.*[2] show that people with tetraplegia can use a neural device, known as BrainGate, to control a robotic arm for reaching and grasping objects. This work builds on decades of previous research on the neural mechanisms that control arm movements[13–15] (blue), on electrode development[16] (orange) and on neural interfaces in monkeys[3–6] (green), which opened the way to studies in humans[1,2] (purple).

in a deterioration of signal quality over time. The authors acknowledge that some deterioration had indeed occurred, but it is encouraging that useful signals could still be obtained five years later. Nevertheless, as many spinal-cord injuries occur at a young age and patients may live with their disabilities for many decades, further efforts to understand and control the tissue response to indwelling electrodes will be crucial for widespread clinical application of neural-interface systems.

The current study also underlines the importance of basic research in driving translational advances. At a time when experimentation using non-human primates is increasingly controversial, it is worth noting that the results reported by Hochberg *et al.* draw directly on previous neural-interface demonstrations in monkeys[3–6] and on decades of basic research into the control of arm movements (Fig. 1). The upper limb of primates is a uniquely versatile tool, and its evolution involved profound changes to the motor structures of the brain and their descending connections that are not shared with other mammals such as mice and rats[7]. Further understanding of how distributed populations of neurons in the brain and the spinal cord cooperate during dexterous manipulation of objects will doubtless inform the development of improved neural interfaces for artificial limbs.

Paralysis resulting from nervous-system injury is a multifaceted condition, with many aspects that affect quality of life. Nevertheless, patients consistently rate regaining arm and hand function as a top priority[8]. So, although robotic arms may be of practical assistance, restoring movements of the patients' own limbs should remain the ultimate goal. Future neural-interface systems may help to achieve this, if they can be coupled to functional electrical stimulation of muscles[9,10] or the spinal cord[11]. In addition, damage to sensory pathways may require artificial sensation to be relayed to the brain before fully naturalistic movements can be restored[12]. Furthermore, all this should preferably be performed by wireless implants that do not physically breach the skin. Encouraging progress is being made on all these fronts.

Ultimately, the greatest obstacle to clinical applications of neural interfaces may come not from science or engineering, but from economics. The original BrainGate clinical trial was initiated by a US company (Cyberkinetics Neurotechnology Systems) that ceased operations in 2009. Fortunately, a new clinical trial is in progress, administered by the Massachusetts General Hospital in Boston. It remains to be seen whether a neural-interface system that will be of practical use to patients with diverse clinical needs can become a commercially viable proposition. Nevertheless, the delight of a participant in Hochberg and colleagues' study as she succeeds in drinking from a bottle for the first time in years (see Supplementary Movie 4 that accompanies the paper[2]) should act as a powerful incentive for all in the field to address these challenges. ■

Andrew Jackson *is at the Institute of Neuroscience, Newcastle University, Newcastle upon Tyne NE2 4HH, UK.*
e-mail: andrew.jackson@ncl.ac.uk

1. Hochberg, L. R. *et al. Nature* **442,** 164–171 (2006).
2. Hochberg, L. R. *et al. Nature* **485,** 372–375 (2012).
3. Wessberg, J. *et al. Nature* **408,** 361–365 (2000).
4. Serruya, M. D., Hatsopoulos, N. G., Paninski, L., Fellows, M. R. & Donoghue, J. P. *Nature* **416,** 141–142 (2002).
5. Taylor, D. M., Tillery, S. I. & Schwartz, A. B. *Science* **296,** 1829–1832 (2002).
6. Velliste, M., Perel, S., Spalding, M. C., Whitford, A. S. & Schwartz, A. B. *Nature* **453,** 1098–1101 (2008).
7. Lemon, R. N. *Annu. Rev. Neurosci.* **31,** 195–218 (2008).
8. Anderson, K. D. *J. Neurotrauma* **21,** 1371–1383 (2004).
9. Moritz, C. T., Perlmutter, S. I. & Fetz, E. E. *Nature* **456,** 639–642 (2008).
10. Ethier, C., Oby, E. R., Bauman, M. J. & Miller, L. E. *Nature* **485,** 368–371 (2012).
11. Zimmermann, J. B., Seki, K. & Jackson, A. *J. Neural Eng.* **8,** 054001 (2011).
12. O'Doherty, J. E. *et al. Nature* **479,** 228–231 (2011).
13. Evarts, E. V. *J. Neurophysiol.* **27,** 152–171 (1964).
14. Humphrey, D. R., Schmidt, E. M. & Thompson W. D. *Science* **170,** 758–762 (1970).
15. Georgopoulos, A. P., Kalaska, J. F., Caminiti, R. & Massey, J. T. *J. Neurosci.* **2,** 1527–1537 (1982).
16. Nordhausen, C. T., Rousche, P. J. & Normann, R. A. *Brain Res.* **637,** 27–36 (1994).

GENETICS

Fish heads and human disease

The expression level of a single gene can determine head size in zebrafish, mirroring a human anatomical feature associated with neurological disorders such as autism and schizophrenia. SEE LETTER P.363

DHEERAJ MALHOTRA & JONATHAN SEBAT

Neurodevelopmental and neuropsychiatric disorders can be caused by a multitude of genetic and environmental factors. The genetic abnormalities associated with disorders such as autism, schizophrenia and early-onset bipolar disorder[1] include a class of mutations called copy-number variants (CNVs), which involve deletion or duplication of whole regions of the genome, typically spanning multiple genes. Among the CNVs most frequently observed in psychiatric disorders is a region of chromosome 16 that contains 29 genes[2,3], called 16p11.2. However, we still have little understanding of the mechanism by which this CNV exerts its clinical effects, and, indeed, which gene or genes are responsible. On page 363 of this issue, Golzio and colleagues show[4] that one of the 16p11.2 CNV genes, *KCTD13*, helps to regulate brain size in zebrafish. This finding provides a tantalizing potential link between *KCTD13* and the abnormalities in brain growth and behaviour that are associated with the 16p11.2 CNV in humans.

At the genetic level, the 16p11.2 CNV comes in two forms: a deletion and a duplication, each of the same 29 genes. The associated clinical presentation of patients can be quite variable, but some human traits have been found to correlate strongly with CNV genotype. Patients with the deletion can show obesity[5,6] and increased head size[7,8], whereas the duplication is associated with leanness[9], decreased head size[7,8] and psychiatric disorders[8]. Both the loss and the gain of 16p11.2 confer a significant risk of autism[10] and developmental delay[2,7].

Because CNVs can span multiple genes, their effects could be due to the loss or gain of either a single gene or a combination of multiple genes, which makes it difficult to study CNVs in animal model systems. A new chromosome-engineered mouse model has faithfully recapitulated the human genotype and some of the human clinical characteristics associated with the 16p11.2 CNV. Most notably, this model has shown[11] that the 16p11.2 CNV has similar effects on head size in humans and mice. But these models have not pinpointed the effects of individual genes.

The success of Golzio and colleagues' strategy can be attributed to their use of the zebrafish as an efficient tool for genetic manipulation and high-throughput screening. Attempting to model psychiatric conditions in fish presents obvious challenges, but the authors overcome some of these difficulties by focusing on the anatomical feature of brain size, which can be readily ascertained in fish.

The researchers show that, in zebrafish, the overexpression of a single gene, *KCTD13*, causes a significant decrease in brain size (microcephaly), whereas inhibition of *KCTD13* expression leads to an increase in brain size (macrocephaly). These changes perfectly mirror the effects of the 16p11.2

in a deterioration of signal quality over time. The authors acknowledge that some deterioration had indeed occurred, but it is encouraging that useful signals could still be obtained five years later. Nevertheless, as many spinal-cord injuries occur at a young age and patients may live with their disabilities for many decades, further efforts to understand and control the tissue response to indwelling electrodes will be crucial for widespread clinical application of neural-interface systems.

The current study also underlines the importance of basic research in driving translational advances. At a time when experimentation using non-human primates is increasingly controversial, it is worth noting that the results reported by Hochberg *et al.* draw directly on previous neural-interface demonstrations in monkeys[3–6] and on decades of basic research into the control of arm movements (Fig. 1). The upper limb of primates is a uniquely versatile tool, and its evolution involved profound changes to the motor structures of the brain and their descending connections that are not shared with other mammals such as mice and rats[7]. Further understanding of how distributed populations of neurons in the brain and the spinal cord cooperate during dexterous manipulation of objects will doubtless inform the development of improved neural interfaces for artificial limbs.

Paralysis resulting from nervous-system injury is a multifaceted condition, with many aspects that affect quality of life. Nevertheless, patients consistently rate regaining arm and hand function as a top priority[8]. So, although robotic arms may be of practical assistance, restoring movements of the patients' own limbs should remain the ultimate goal. Future neural-interface systems may help to achieve this, if they can be coupled to functional electrical stimulation of muscles[9,10] or the spinal cord[11]. In addition, damage to sensory pathways may require artificial sensation to be relayed to the brain before fully naturalistic movements can be restored[12]. Furthermore, all this should preferably be performed by wireless implants that do not physically breach the skin. Encouraging progress is being made on all these fronts.

Ultimately, the greatest obstacle to clinical applications of neural interfaces may come not from science or engineering, but from economics. The original BrainGate clinical trial was initiated by a US company (Cyberkinetics Neurotechnology Systems) that ceased operations in 2009. Fortunately, a new clinical trial is in progress, administered by the Massachusetts General Hospital in Boston. It remains to be seen whether a neural-interface system that will be of practical use to patients with diverse clinical needs can become a commercially viable proposition. Nevertheless, the delight of a participant in Hochberg and colleagues' study as she succeeds in drinking from a bottle for the first time in years (see Supplementary Movie 4 that accompanies the paper[2]) should act as a powerful incentive for all in the field to address these challenges. ■

Andrew Jackson *is at the Institute of Neuroscience, Newcastle University, Newcastle upon Tyne NE2 4HH, UK.*
e-mail: andrew.jackson@ncl.ac.uk

1. Hochberg, L. R. *et al. Nature* **442,** 164–171 (2006).
2. Hochberg, L. R. *et al. Nature* **485,** 372–375 (2012).
3. Wessberg, J. *et al. Nature* **408,** 361–365 (2000).
4. Serruya, M. D., Hatsopoulos, N. G., Paninski, L., Fellows, M. R. & Donoghue, J. P. *Nature* **416,** 141–142 (2002).
5. Taylor, D. M., Tillery, S. I. & Schwartz, A. B. *Science* **296,** 1829–1832 (2002).
6. Velliste, M., Perel, S., Spalding, M. C., Whitford, A. S. & Schwartz, A. B. *Nature* **453,** 1098–1101 (2008).
7. Lemon, R. N. *Annu. Rev. Neurosci.* **31,** 195–218 (2008).
8. Anderson, K. D. *J. Neurotrauma* **21,** 1371–1383 (2004).
9. Moritz, C. T., Perlmutter, S. I. & Fetz, E. E. *Nature* **456,** 639–642 (2008).
10. Ethier, C., Oby, E. R., Bauman, M. J. & Miller, L. E. *Nature* **485,** 368–371 (2012).
11. Zimmermann, J. B., Seki, K. & Jackson, A. *J. Neural Eng.* **8,** 054001 (2011).
12. O'Doherty, J. E. *et al. Nature* **479,** 228–231 (2011).
13. Evarts, E. V. *J. Neurophysiol.* **27,** 152–171 (1964).
14. Humphrey, D. R., Schmidt, E. M. & Thompson W. D. *Science* **170,** 758–762 (1970).
15. Georgopoulos, A. P., Kalaska, J. F., Caminiti, R. & Massey, J. T. *J. Neurosci.* **2,** 1527–1537 (1982).
16. Nordhausen, C. T., Rousche, P. J. & Normann, R. A. *Brain Res.* **637,** 27–36 (1994).

GENETICS

Fish heads and human disease

The expression level of a single gene can determine head size in zebrafish, mirroring a human anatomical feature associated with neurological disorders such as autism and schizophrenia. SEE LETTER P.363

DHEERAJ MALHOTRA & JONATHAN SEBAT

Neurodevelopmental and neuropsychiatric disorders can be caused by a multitude of genetic and environmental factors. The genetic abnormalities associated with disorders such as autism, schizophrenia and early-onset bipolar disorder[1] include a class of mutations called copy-number variants (CNVs), which involve deletion or duplication of whole regions of the genome, typically spanning multiple genes. Among the CNVs most frequently observed in psychiatric disorders is a region of chromosome 16 that contains 29 genes[2,3], called 16p11.2. However, we still have little understanding of the mechanism by which this CNV exerts its clinical effects, and, indeed, which gene or genes are responsible. On page 363 of this issue, Golzio and colleagues show[4] that one of the 16p11.2 CNV genes, *KCTD13*, helps to regulate brain size in zebrafish. This finding provides a tantalizing potential link between *KCTD13* and the abnormalities in brain growth and behaviour that are associated with the 16p11.2 CNV in humans.

At the genetic level, the 16p11.2 CNV comes in two forms: a deletion and a duplication, each of the same 29 genes. The associated clinical presentation of patients can be quite variable, but some human traits have been found to correlate strongly with CNV genotype. Patients with the deletion can show obesity[5,6] and increased head size[7,8], whereas the duplication is associated with leanness[9], decreased head size[7,8] and psychiatric disorders[8]. Both the loss and the gain of 16p11.2 confer a significant risk of autism[10] and developmental delay[2,7].

Because CNVs can span multiple genes, their effects could be due to the loss or gain of either a single gene or a combination of multiple genes, which makes it difficult to study CNVs in animal model systems. A new chromosome-engineered mouse model has faithfully recapitulated the human genotype and some of the human clinical characteristics associated with the 16p11.2 CNV. Most notably, this model has shown[11] that the 16p11.2 CNV has similar effects on head size in humans and mice. But these models have not pinpointed the effects of individual genes.

The success of Golzio and colleagues' strategy can be attributed to their use of the zebrafish as an efficient tool for genetic manipulation and high-throughput screening. Attempting to model psychiatric conditions in fish presents obvious challenges, but the authors overcome some of these difficulties by focusing on the anatomical feature of brain size, which can be readily ascertained in fish.

The researchers show that, in zebrafish, the overexpression of a single gene, *KCTD13*, causes a significant decrease in brain size (microcephaly), whereas inhibition of *KCTD13* expression leads to an increase in brain size (macrocephaly). These changes perfectly mirror the effects of the 16p11.2

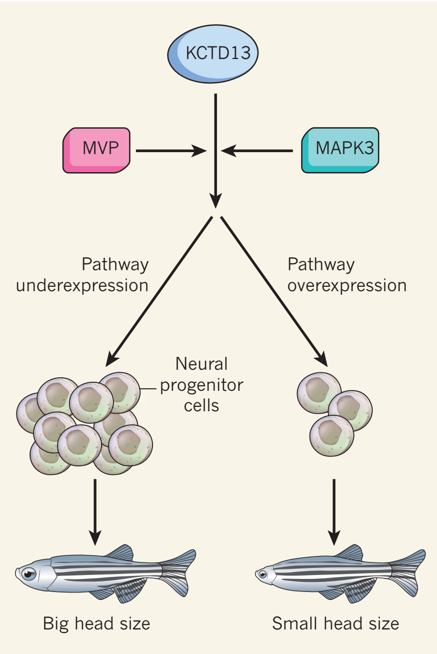


Figure 1 | A gene for head size. Golzio *et al.*[4] show that overexpression of a single gene, *KCTD13*, causes abnormally small head size in zebrafish embryos, and that inhibition of the gene leads to oversized heads. This effect apparently arises from the KCTD13 protein's influence on the growth of neural progenitor cells — the protein's presence leads to cell death, but in its absence cell proliferation is enhanced. The authors also show that the expression level of two of *KCTD13*'s neighbouring genes, *MVP* and *MAPK3*, enhance the effect of *KCTD13* on fish head size. All three genes belong to the 16p11.2 copy-number variant, a mutation that is associated with abnormal brain growth and neurodevelopmental disorders in humans.

CNV on head size in humans and mice[8,11].

KCTD13 expression levels seem to influence brain growth by regulating neuron cell number during development. When Golzio and colleagues measured the rates of proliferation and death among neuronal progenitor cells in zebrafish and mice, they found that the *KCTD13* dosage modulates early neurogenesis, or neuron formation. Increased *KCTD13* expression induces programmed cell death of neuronal progenitors, whereas decreased expression leads to increased progenitor-cell proliferation (Fig. 1).

These findings suggest that *KCTD13* is responsible for the abnormalities in brain growth associated with 16p11.2 copy number. However, *KCTD13* may not be the only gene involved. By performing pairwise overexpressions of *KCTD13* with each of the remaining 16p11.2 genes, the authors show that two other genes — *MAPK3* and *MVP* — interact with *KCTD13* to increase its influence on brain size.

Is *KCTD13* a key gene for autism risk in humans? Although plausible, this possibility has not been confirmed, and genetic evidence from humans is still largely anecdotal. This study and a previous one[12] identified two individuals with autism who had small deletions of *KCTD13*, but confirming the role of *KCTD13* in autism and related disorders will require additional studies in humans and in mice.

The function of the protein encoded by *KCTD13* is unclear. It is known[13] that KCTD13 interacts directly with a protein called proliferating cell nuclear antigen (PCNA), which is involved in numerous cell processes including DNA replication and repair, and the assembly of chromatin — the DNA–protein complex that makes up the chromosome. This activity is consistent with KCTD13 having a role in the regulation of the cell cycle. Intriguingly, the proteins encoded by *MAPK3* and *MVP* — the two genes that were found to interact with *KCTD13* — also regulate cell proliferation[14,15]. Thus, Golzio and colleagues' findings allow for a speculative but coherent model of the molecular, cellular and neuroanatomical mechanism of disease in autism: changes in the number of copies of *KCTD13*, *MAPK3* and *MVP* directly affect cell-cycle regulation and cell proliferation, thereby leading to abnormal brain growth. This notion is supported by the existence of other brain-overgrowth syndromes in humans that are also associated with mutations in genes that control cell-cycle progression[16], including tuberous sclerosis, neurofibromatosis, Cowden syndrome and Sotos syndrome.

Golzio and colleagues' use of the zebrafish, the little aquarium fish originally from the River Ganges, has provided a big clue to the molecular and cellular mechanisms underlying the conditions associated with the 16p11.2 CNV. Similar studies might prove effective in elucidating the relationship between genes and neurodevelopment in other genetic disorders and in complex inherited disease. ■

Dheeraj Malhotra *and* **Jonathan Sebat** *are in the Departments of Psychiatry, and of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, California 92093, USA.* **J.S.** *is also at the Institute for Genomic Medicine and the Beyster Center for Genomics of Psychiatric Diseases, University of California, San Diego.*
e-mail: jsebat@ucsd.edu

1. Malhotra, D. & Sebat, J. *Cell* **148,** 1223–1241 (2012).
2. Cooper, G. M. *et al. Nature Genet.* **43,** 838–846 (2011).
3. Kaminsky, E. B. *et al. Genet. Med.* **13,** 777–784 (2011).
4. Golzio, C. *et al. Nature* **485,** 363–367 (2012).
5. Bochukova, E. G. *et al. Nature* **463,** 666–670 (2010).
6. Walters, R. G. *et al. Nature* **463,** 671–675 (2010).
7. Shinawi, M. *et al. J. Med. Genet.* **47,** 332–341 (2010).
8. McCarthy, S. E. *et al. Nature Genet.* **41,** 1223–1227 (2009).
9. Jacquemont, S. *et al. Nature* **478,** 97–102 (2011).
10. Weiss, L. A. *et al. N. Engl. J. Med.* **358,** 667–675 (2008).
11. Horev, G. *et al. Proc. Natl Acad. Sci. USA* **108,** 17076–17081 (2011).
12. Crepel, A. *et al. Am. J. Med. Genet. B* **156,** 243–245 (2011).
13. He, H., Tan, C. K., Downey, K. M. & So, A. G. *Proc. Natl Acad. Sci. USA* **98,** 11979–11984 (2001).
14. Scheffer, G. L. *et al. Nature Med.* **1,** 578–582 (1995).
15. Zhang, W. & Liu, H. T. *Cell Res.* **12,** 9–18 (2002).
16. Cohen, M. M. Jr *Am. J. Med. Genet. C* **117,** 49–56 (2003).

EARTH SCIENCE

Geomagnetism under scrutiny

New calculations show that the electrical resistance of Earth's liquid-iron core is lower than had been thought. The results prompt a reassessment of how the planet's magnetic field has been generated and maintained over time. SEE LETTER P.355

BRUCE BUFFETT

Fluid flow in Earth's liquid-iron core sustains the planet's magnetic field against persistent losses due to the electrical resistance of liquid iron. The battle between generation and loss of the magnetic field suggests that a decrease in electrical resistance would tilt the balance in favour of generation. Surprisingly, the opposite may be closer to the truth. On page 355 of this issue, Pozzo *et al.*[1] use a mathematical method called density functional theory to predict the electrical resistance of liquid-iron alloys at the high pressures and temperatures found in Earth's core. Their values are only two to three times lower than previous estimates[2], but this change is large enough to affect our understanding of the dynamics and evolution of Earth's interior.

A good electrical conductor is essential for generating a planetary magnetic field. The movement of conducting material induces electric currents, which reinforce the initial magnetic field. A self-sustaining process requires that the effect of fluid motion exceeds the loss due to the conductor's electrical resistance. The ratio of these

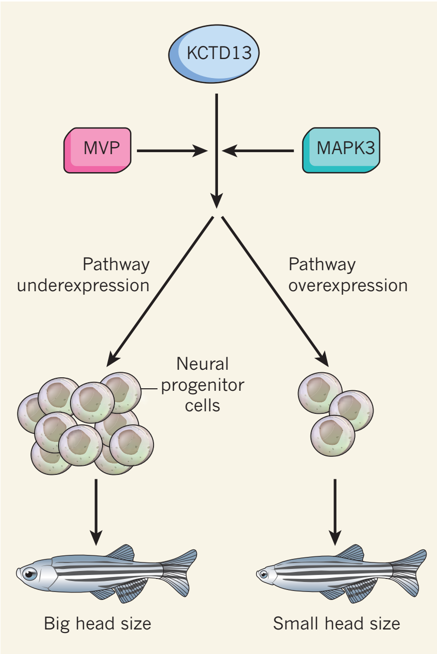


Figure 1 | A gene for head size. Golzio *et al.*[4] show that overexpression of a single gene, *KCTD13*, causes abnormally small head size in zebrafish embryos, and that inhibition of the gene leads to oversized heads. This effect apparently arises from the KCTD13 protein's influence on the growth of neural progenitor cells — the protein's presence leads to cell death, but in its absence cell proliferation is enhanced. The authors also show that the expression level of two of *KCTD13*'s neighbouring genes, *MVP* and *MAPK3*, enhance the effect of *KCTD13* on fish head size. All three genes belong to the 16p11.2 copy-number variant, a mutation that is associated with abnormal brain growth and neurodevelopmental disorders in humans.

CNV on head size in humans and mice[8,11].

KCTD13 expression levels seem to influence brain growth by regulating neuron cell number during development. When Golzio and colleagues measured the rates of proliferation and death among neuronal progenitor cells in zebrafish and mice, they found that the *KCTD13* dosage modulates early neurogenesis, or neuron formation. Increased *KCTD13* expression induces programmed cell death of neuronal progenitors, whereas decreased expression leads to increased progenitor-cell proliferation (Fig. 1).

These findings suggest that *KCTD13* is responsible for the abnormalities in brain growth associated with 16p11.2 copy number. However, *KCTD13* may not be the only gene involved. By performing pairwise overexpressions of *KCTD13* with each of the remaining 16p11.2 genes, the authors show that two other genes — *MAPK3* and *MVP* — interact with *KCTD13* to increase its influence on brain size.

Is *KCTD13* a key gene for autism risk in humans? Although plausible, this possibility has not been confirmed, and genetic evidence from humans is still largely anecdotal. This study and a previous one[12] identified two individuals with autism who had small deletions of *KCTD13*, but confirming the role of *KCTD13* in autism and related disorders will require additional studies in humans and in mice.

The function of the protein encoded by *KCTD13* is unclear. It is known[13] that KCTD13 interacts directly with a protein called proliferating cell nuclear antigen (PCNA), which is involved in numerous cell processes including DNA replication and repair, and the assembly of chromatin — the DNA–protein complex that makes up the chromosome. This activity is consistent with KCTD13 having a role in the regulation of the cell cycle. Intriguingly, the proteins encoded by *MAPK3* and *MVP* — the two genes that were found to interact with *KCTD13* — also regulate cell proliferation[14,15]. Thus, Golzio and colleagues' findings allow for a speculative but coherent model of the molecular, cellular and neuroanatomical mechanism of disease in autism: changes in the number of copies of *KCTD13*, *MAPK3* and *MVP* directly affect cell-cycle regulation and cell proliferation, thereby leading to abnormal brain growth. This notion is supported by the existence of other brain-overgrowth syndromes in humans that are also associated with mutations in genes that control cell-cycle progression[16], including tuberous sclerosis, neurofibromatosis, Cowden syndrome and Sotos syndrome.

Golzio and colleagues' use of the zebrafish, the little aquarium fish originally from the River Ganges, has provided a big clue to the molecular and cellular mechanisms underlying the conditions associated with the 16p11.2 CNV. Similar studies might prove effective in elucidating the relationship between genes and neurodevelopment in other genetic disorders and in complex inherited disease. ■

Dheeraj Malhotra *and* **Jonathan Sebat** *are in the Departments of Psychiatry, and of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, California 92093, USA.* **J.S.** *is also at the Institute for Genomic Medicine and the Beyster Center for Genomics of Psychiatric Diseases, University of California, San Diego.*
e-mail: jsebat@ucsd.edu

1. Malhotra, D. & Sebat, J. *Cell* **148,** 1223–1241 (2012).
2. Cooper, G. M. *et al. Nature Genet.* **43,** 838–846 (2011).
3. Kaminsky, E. B. *et al. Genet. Med.* **13,** 777–784 (2011).
4. Golzio, C. *et al. Nature* **485,** 363–367 (2012).
5. Bochukova, E. G. *et al. Nature* **463,** 666–670 (2010).
6. Walters, R. G. *et al. Nature* **463,** 671–675 (2010).
7. Shinawi, M. *et al. J. Med. Genet.* **47,** 332–341 (2010).
8. McCarthy, S. E. *et al. Nature Genet.* **41,** 1223–1227 (2009).
9. Jacquemont, S. *et al. Nature* **478,** 97–102 (2011).
10. Weiss, L. A. *et al. N. Engl. J. Med.* **358,** 667–675 (2008).
11. Horev, G. *et al. Proc. Natl Acad. Sci. USA* **108,** 17076–17081 (2011).
12. Crepel, A. *et al. Am. J. Med. Genet. B* **156,** 243–245 (2011).
13. He, H., Tan, C. K., Downey, K. M. & So, A. G. *Proc. Natl Acad. Sci. USA* **98,** 11979–11984 (2001).
14. Scheffer, G. L. *et al. Nature Med.* **1,** 578–582 (1995).
15. Zhang, W. & Liu, H. T. *Cell Res.* **12,** 9–18 (2002).
16. Cohen, M. M. Jr *Am. J. Med. Genet. C* **117,** 49–56 (2003).

EARTH SCIENCE

Geomagnetism under scrutiny

New calculations show that the electrical resistance of Earth's liquid-iron core is lower than had been thought. The results prompt a reassessment of how the planet's magnetic field has been generated and maintained over time. SEE LETTER P.355

BRUCE BUFFETT

Fluid flow in Earth's liquid-iron core sustains the planet's magnetic field against persistent losses due to the electrical resistance of liquid iron. The battle between generation and loss of the magnetic field suggests that a decrease in electrical resistance would tilt the balance in favour of generation. Surprisingly, the opposite may be closer to the truth. On page 355 of this issue, Pozzo *et al.*[1] use a mathematical method called density functional theory to predict the electrical resistance of liquid-iron alloys at the high pressures and temperatures found in Earth's core. Their values are only two to three times lower than previous estimates[2], but this change is large enough to affect our understanding of the dynamics and evolution of Earth's interior.

A good electrical conductor is essential for generating a planetary magnetic field. The movement of conducting material induces electric currents, which reinforce the initial magnetic field. A self-sustaining process requires that the effect of fluid motion exceeds the loss due to the conductor's electrical resistance. The ratio of these

two effects is often characterized by the magnetic Reynolds number, which must exceed a threshold value for a magnetic field to be sustained[3]. High fluid velocity and/or low electrical resistance promote a large Reynolds number. It follows that a low resistance should enhance field generation, but only if the velocity is maintained at the necessary level.

Metals are good thermal conductors because electrons are more effective than atomic vibrations in transporting heat. Pozzo and colleagues' simulations confirm that the thermal conductivity of liquid iron under the conditions in Earth's core is several times higher than previous estimates[2]. They predict a value of roughly 125 watts per metre per kelvin ($W\,m^{-1}\,K^{-1}$) at the top of the core and more than 200 $W\,m^{-1}\,K^{-1}$ at the boundary between the outer and inner parts of the core. Such large thermal conductivities allow a substantial amount of heat to be carried by conduction, leaving less heat to drive convection. Convection may even cease in parts of the core[4].

To illustrate the situation, let us consider a representative temperature profile in the liquid core (Fig. 1). Increase of temperature with depth (or pressure) causes conduction of heat towards the top of the core. The depth dependence of temperature in a convecting fluid is well approximated by an adiabatic profile, which is based on the idea that rising and sinking parcels of fluid do not exchange heat. When Pozzo *et al.* applied their new estimate for the thermal conductivity to an adiabatic profile in the core, they obtained a conductive heat flow of 15 terawatts (10^{12} W) near the top of the core. This value may exceed the heat flow across the core–mantle boundary[5]. In that case, warm fluid would accumulate at the top of the core, creating a stably stratified layer. As a result, convection and magnetic-field generation would be largely confined to the region below the stratified layer.

Pozzo *et al.* assess the consequences of high thermal conductivity for magnetic-field generation by constructing thermal 'histories' for the core. They present a suite of histories that could sustain a magnetic field, but in each case a very thick stratified layer or an additional energy source due to decay of radioactive elements would be required in the core. Reasonable arguments can be made against both of these options[6,7], but one or the other seems to be unavoidable for maintaining the magnetic field. If Pozzo and colleagues' calculations are correct, then some of our basic assumptions about the core must be wrong.

One might question the calculations that predict high thermal conductivities. However, similar results have been obtained in independent calculations[8], and there is further experimental support[9] (albeit at temperatures much lower than core conditions). Accepting high thermal conductivity means that heat loss through conduction would substantially weaken thermal convection. A modest (subadiabatic) heat flow from the core would confine convection to a small region below a thick, thermally stratified layer. Such a layer would suppress variations in the magnetic field with time, which is at odds with observations. Alternatively, the need for a stratified layer could be eliminated if the heat flow from the core exceeded the heat conducted along the adiabatic profile. However, it is unclear how this high heat flow could be maintained over geological time. Indeed, most studies suggest that the heat flow from the core was higher in the past[10]. Perhaps the answer involves an unknown energy source. For example, chemical interactions between the core and the mantle might draw on the planet's gravitational energy. However, lack of the necessary understanding of the relevant chemistry at high pressures and temperatures means that this possibility cannot be assessed.

A high thermal conductivity for the liquid-iron core also has implications for the dynamics of the solid inner core. Iron at inner-core conditions is under higher pressure, and probably has a lower concentration of impurities, than iron in the overlying liquid core. Both of these factors would increase the thermal conductivity, so the value in the inner core should exceed 200 $W\,m^{-1}\,K^{-1}$. Such a high value would make convection in the inner core[11], including its 'translational' form[12,13], unlikely. Instead, the inner core should cool by conduction. In such conditions, strong thermal stratification would develop and radial motion would effectively be suppressed. Because radial motion in the inner core is often invoked to explain the directional dependence (anisotropy) of seismic-wave speed in the inner core[14], the high values of thermal conductivity should force researchers to look elsewhere for the cause of the seismic anisotropy.

It is remarkable that a modest change in thermal conductivity can have such a dramatic affect on the dynamics of Earth's core. More broadly, the latest study reveals how the properties of liquid iron make the operation of magnetic dynamos in terrestrial planets even more precarious than was previously believed. We are left with the challenge of understanding how Earth has succeeded in maintaining its magnetic field over most of geological time. ■

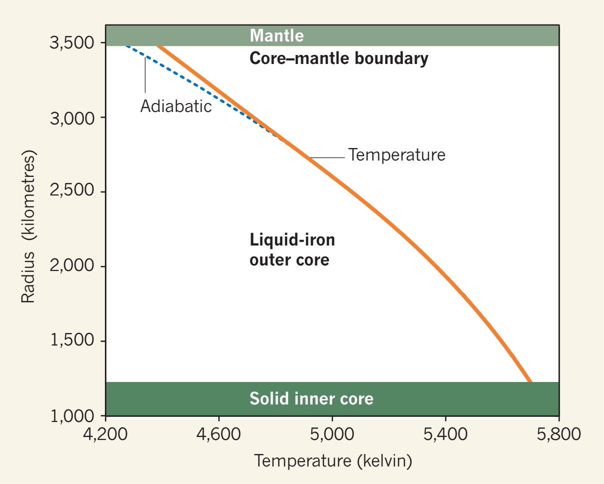


Figure 1 | Temperature profile of Earth's interior. The liquid-iron outer core lies between the mantle and the solid inner core. The increase in temperature across the liquid outer core is well described by an adiabatic process, in which no heat transfer occurs between ascending and descending liquid-iron parcels. Thermal conduction carries heat towards the top of the core. Pozzo and colleagues' study[1] indicates that heat conduction inside the core may exceed the flow of heat across the core–mantle boundary. As a result, the temperature in the boundary region departs from the adiabatic profile to match the boundary's heat flow.

Bruce Buffett *is in the Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, California 94720-4767, USA.*
e-mail: bbuffett@berkeley.edu

1. Pozzo, M., Davies, C., Gubbins, D. & Alfè, D. *Nature* **485,** 355–358 (2012).
2. Stacey, F. D. & Anderson, O. L. *Phys. Earth Planet. Inter.* **124,** 153–162 (2001).
3. Christensen, U. R. & Aubert, J. *Geophys. J. Int.* **166,** 97–114 (2006).
4. Gubbins, D., Thomson, C. J. & Whaler, K. A. *Geophys. J. R. Astron. Soc.* **68,** 241–251 (1982).
5. Lay, T., Hernlund, J. & Buffett, B. A. *Nature Geosci.* **1,** 25–32 (2008).
6. Gillet, N., Schaeffer, N. & Jault, D. *Phys. Earth Planet. Inter.* **187,** 380–390 (2011).
7. Corgne, A., Keshav, S., Fei, Y. W. & McDonough, W. F. *Earth Planet. Sci. Lett.* **256,** 567–576 (2007).
8. de Koker, N., Steinle-Neumann, G. & Vlček, V. *Proc. Natl Acad. Sci. USA* **109,** 4070–4073 (2012).
9. Hirose, K. *et al. Mineral. Mag.* **75,** 1027 (2011).
10. Nakagawa, T. & Tackley, P. J. *Geochem. Geophys. Geosyst.* **11,** Q06001 (2010).
11. Buffett, B. A. *Geophys. J. Int.* **179,** 711–719 (2009).
12. Monnereau, M. *et al. Science* **328,** 1014–1017 (2010).
13. Alboussière, T., Deguen, R. & Melzani, M. *Nature* **466,** 744–747 (2010).
14. Sun, X. & Song, X. *Phys. Earth Planet. Inter.* **167,** 53–70 (2008).

CORRECTION

In the News & Views article 'Cancer biology: The director's cut' by Antonio Gentilella and George Thomas (*Nature* **485,** 50–51; 2012), the messenger RNA transcript encoding YB1 was incorrectly referred to as a 5′ TOP mRNA. The transcript should have been described as containing a pyrimidine-rich translational element (PRTE).

doi:10.1038/nature11040

Cardiac angiogenic imbalance leads to peripartum cardiomyopathy

Ian S. Patten[1,2]*, Sarosh Rana[3]*, Sajid Shahul[4], Glenn C. Rowe[1], Cholsoon Jang[1], Laura Liu[1], Michele R. Hacker[3], Julie S. Rhee[3], John Mitchell[4], Feroze Mahmood[4], Philip Hess[4], Caitlin Farrell[1], Nicole Koulisis[1], Eliyahu V. Khankin[5], Suzanne D. Burke[5,8], Igor Tudorache[6], Johann Bauersachs[7], Federica del Monte[1], Denise Hilfiker-Kleiner[7], S. Ananth Karumanchi[5,8] & Zoltan Arany[1]

Peripartum cardiomyopathy (PPCM) is an often fatal disease that affects pregnant women who are near delivery, and it occurs more frequently in women with pre-eclampsia and/or multiple gestation. The aetiology of PPCM, and why it is associated with pre-eclampsia, remain unknown. Here we show that PPCM is associated with a systemic angiogenic imbalance, accentuated by pre-eclampsia. Mice that lack cardiac PGC-1α, a powerful regulator of angiogenesis, develop profound PPCM. Importantly, the PPCM is entirely rescued by pro-angiogenic therapies. In humans, the placenta in late gestation secretes VEGF inhibitors like soluble FLT1 (sFLT1), and this is accentuated by multiple gestation and pre-eclampsia. This anti-angiogenic environment is accompanied by subclinical cardiac dysfunction, the extent of which correlates with circulating levels of sFLT1. Exogenous sFLT1 alone caused diastolic dysfunction in wild-type mice, and profound systolic dysfunction in mice lacking cardiac PGC-1α. Finally, plasma samples from women with PPCM contained abnormally high levels of sFLT1. These data indicate that PPCM is mainly a vascular disease, caused by excess anti-angiogenic signalling in the peripartum period. The data also explain how late pregnancy poses a threat to cardiac homeostasis, and why pre-eclampsia and multiple gestation are important risk factors for the development of PPCM.

PPCM affects 1 in 300 to 1 in 3,000 pregnancies, with geographic hot spots of high incidence, such as Nigeria and Haiti[1,2]. The disease is characterized by systolic heart failure presenting in the last month of pregnancy or the first 5 months post-partum. Although approximately half of affected women recover cardiac function post-partum, many patients progress to chronic heart failure, cardiac transplantation or death. Thus, PPCM can devastate otherwise healthy young women and their infants. PPCM remains a disease of unknown aetiology. The onset late in gestation does not coincide with increased haemodynamic load on the heart, suggesting that other mechanisms are responsible. Recent data have suggested that anti-angiogenic prolactin fragments may have an important role in causing the disease in some patients[3]. Risk factors for PPCM also include pre-eclampsia and multiple gestation, suggesting potential mechanistic overlap with these processes[1,2].

PGC-1α is a transcriptional coactivator that drives mitochondrial biogenesis and other metabolic programs in many tissues, including the heart[4,5]. PGC-1α is highly expressed in the heart, and mice lacking PGC-1α globally have abnormal cardiac energetic reserves and respond poorly to stressful stimuli such as transverse aortic banding[6,7]. In addition to its role in mitochondrial homeostasis, PGC-1α also induces the expression and secretion of pro-angiogenic factors, such as vascular endothelial growth factor (VEGF), which leads to formation of new blood vessels[8,9]. Although the angiogenic function of PGC-1α has been described in skeletal muscle, its role in cardiac tissue remains unexplored.

Cardiac-specific PGC-1α deletion leads to PPCM

To study further the role of PGC-1α in the heart, we generated cardiac-specific PGC-1α knockout (HKO) mice (see Methods). While studying these mice, we noticed that female HKO mice were fertile, and delivered normal litter sizes (not shown), but invariably died after one or two pregnancies (Fig. 1a). The hearts of these mice were large, dilated and fibrotic (Fig. 1b–d), consistent with a dilated cardiomyopathy. Two-dimensional M-mode echocardiography revealed dilated, poorly contractile hearts in HKO mice after their second delivery (Fig. 1e). Left ventricular end-diastolic dimensions (LVEDD) and left ventricular end-systolic dimensions (LVESD) were markedly enlarged, and fractional shortening, a direct measure of cardiac contractile function, was profoundly depressed (Fig. 1f–i). Nulliparous mice, as well as post-partum control mice, were not affected. Males were also not affected (Supplementary Fig. 1). Thus, the absence of PGC-1α in cardiomyocytes leads to a profound PPCM in mice.

PGC-1α regulates angiogenesis in cardiac tissue

We have recently shown in skeletal muscle that PGC-1α regulates angiogenesis by driving the expression of angiogenic factors like VEGF[8,9]. Anti-angiogenic therapies, including antibodies that neutralize VEGF and small-molecule VEGF receptor inhibitors, are increasingly being used in oncological and ophthalmological treatments, and cardiomyopathy and heart failure have recently been recognized as important side effects[10,11], showing that anti-angiogenic therapy can be harmful to the heart in humans. Impaired VEGF signalling has also been linked with cardiac dysfunction in mice[12,13]. At the same time, late pregnancy is a strong anti-angiogenic environment, partly owing to the secretion by the placenta of anti-angiogenic factors, like sFLT1, that bind to and neutralize soluble members of the VEGF family[14]. These observations led us to postulate that PGC-1α regulates an angiogenic

[1]Cardiovascular Institute, Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Avenue, Boston, Massachusetts 02115, USA. [2]Center for Vascular Biology Research, Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Avenue, Boston, Massachusetts 02115, USA. [3]Division of Maternal Fetal Medicine/Department of Obstetrics and Gynecology, Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Avenue, Boston, Massachusetts 02115, USA. [4]Department of Anesthesia and Critical Care, Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Avenue, Boston, Massachusetts 02115, USA. [5]Division of Nephrology/Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, 330 Brookline Avenue, Boston, Massachusetts 02115, USA. [6]Department of Cardiothoracic, Transplantation and Vascular Surgery, Medizinische Hochschule Hannover, Carl-Neuberg Strasse, D-30625 Hannover, Germany. [7]Department of Cardiology and Angiology, MedizinischeHochschule Hannover, Carl-Neuberg Strasse, D-30625 Hannover, Germany. [8]Howard Hughes Medical Institute, 330 Brookline Avenue, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

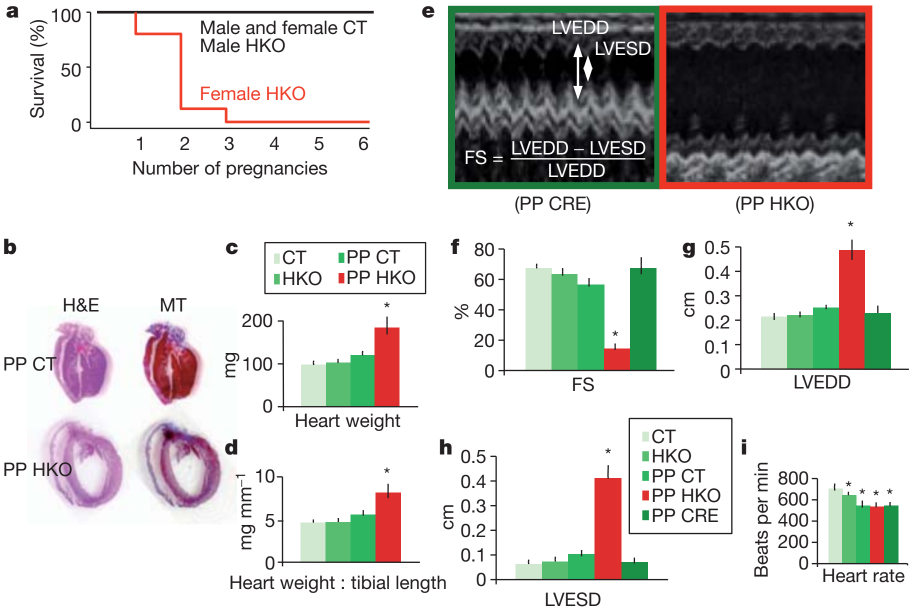


Figure 1 | Mice lacking cardiac PGC-1α develop peri-partum cardiomyopathy. **a**, Kaplan–Meier survival curve in female αMHC-Cre: PGC-$1\alpha^{lox/lox}$ mice (HKO) versus male HKO mice or control mice (CT) of either gender. **b**, Haematoxylin and eosin and Masson's trichrome stains of hearts from post-partum HKO mice (PP HKO), versus post-partum CT mice (PP CT). **c**, **d**, Heart weight (**c**) and heart weight:tibial length ratios (**d**) of nulliparous CT and HKO mice, and PP CT and HKO mice. **e**, Sample M-mode echocardiograms of PP HKO mice and control mice containing the αMHC–Cre transgene alone (PP CRE). **f–i**, Echocardiographic measures in mice of the indicated genotypes, either nulliparous or post-partum. $n \geq 5$ for all groups. $^*P < 0.05$.

program in cardiomyocytes, and that the absence of this program would leave the hearts defenceless to the anti-angiogenic setting of late pregnancy, thus leading the animals to develop PPCM.

Consistent with this idea, overexpression of PGC-1α in neonatal rat ventricular myocytes (NRVMs) strongly induced angiogenic genes involved in the activation and recruitment of endothelial cells (for example, *Vegfa*) and mural cells (for example, *Pdgfb*), as well as genes that are involved in the mitochondrial respiratory chain[15] (for example, *Cycs* and *Cox5b*) (Fig. 2a). Interestingly, some angiogenic genes were repressed (for example, *Bfgf*), with a pattern of repression that is similar to that seen in skeletal muscle cells[8], indicating that PGC-1α reprograms the angiogenic program in a stereotypical manner. Conversely, PGC-1α short interfering RNA (siPGC-1α) significantly suppressed the expression of *Vegfa* (Fig. 2b). Endothelial activation and migration is a hallmark of angiogenesis. As shown in Fig. 2c, d, overexpression of PGC-1α in NRVMs led to a marked, dose-dependent, up to sixfold increase in the migration of the adjacent human umbilical vein endothelial cells (HUVECs) in a co-culture system. Addition of sFLT1 completely neutralized the induced endothelial migration, indicating that secreted members of the VEGF family, probably VEGFA itself, are critical for the effect. Conversely, repression of PGC-1α in NRVMs by siRNA significantly repressed endothelial migration (Fig. 2e). Thus, PGC-1α controls an angiogenic program in cardiomyocytes.

To test whether PGC-1α is required for this program in intact animals, levels of *Vegfa* and other angiogenic factors were measured in hearts from PGC-1α HKO mice. The expression of *Vegfa* and *Pdgfb*, and a number of other angiogenic factors, was repressed in these hearts by as much as 50% (Fig. 3a). VEGFA protein was decreased by 30% in HKO hearts (Fig. 3b). Levels of VEGFA are normally tightly regulated, and even global haplo-insufficiency is lethal[16,17], underscoring the significance of a 30% drop in protein levels. Consistent with these findings, the capillary density of HKO hearts, as measured by staining with the endothelial-specific marker CD31, was decreased by about 15% (Fig. 3c, d). Thus, PGC-1α regulates both vascular density (these data) and mitochondrial function[18] in the heart, providing an important regulatory link between the delivery of fuel (through blood vessels) and its consumption (by mitochondria).These data suggest that PPCM in these mice might be caused by the combination of a heart-specific vascular defect caused by the absence of PGC-1α, and the normal systemic anti-angiogenic environment of late pregnancy. Indeed, the vascular density in HKO hearts decreased by almost 50% post-partum (Fig. 3c, d), which is equivalent to the decrease in vascular density seen in mice lacking cardiac VEGF[19]. Consistent with this vascular rarefaction, perfusion of post-partum HKO hearts was profoundly decreased by nearly 50% compared to wild-type animals, as determined by methoxyisobutylisonitrile (MIBI) uptake and single-photon emission computed tomography and X-ray computed tomography (SPECT/CT) (Fig. 3e, f).

Pro-angiogenic therapy rescues PPCM

To test directly the idea that an angiogenic imbalance drives PPCM in HKO mice, rescue experiments were carried out. In a first series of experiments, breeding mice were administered daily subcutaneous injections of VEGF121 protein ($100\,\mu g\,kg^{-1}$), an isoform of VEGFA, versus vehicle control (Supplementary Fig. 2). The efficacy of VEGF121 injections was confirmed by the presence of VEGF121 in the serum, and robust activation of cardiac VEGFR2 phosphorylation within 30 min of injection (Supplementary Fig. 2b, c). The VEGF treatment led to improved survival of the breeding HKO females; they were able to survive up to five pregnancies (Supplementary Fig. 2d). However, capillary density was only partly rescued, cardiac contractility was only marginally improved and the hearts remained enlarged (Supplementary Fig. 2e–g). Thus, VEGF administration partly rescued lethality in multiparous HKO animals, but it was insufficient to fully rescue PPCM, suggesting that other pathways are also important.

It has recently been shown that STAT3 regulates mitochondrial superoxide dismutase (MNSOD) and protects against reactive oxygen species (ROS) in the heart[3]. Absence of cardiac STAT3 and the consequent increase in ROS led to inappropriate cleavage of prolactin to a potent anti-angiogenic 16-kDa form, and subsequent PPCM (Fig. 4a). Importantly, the PPCM could be rescued by treatment with bromocriptine, which inhibits the secretion of prolactin

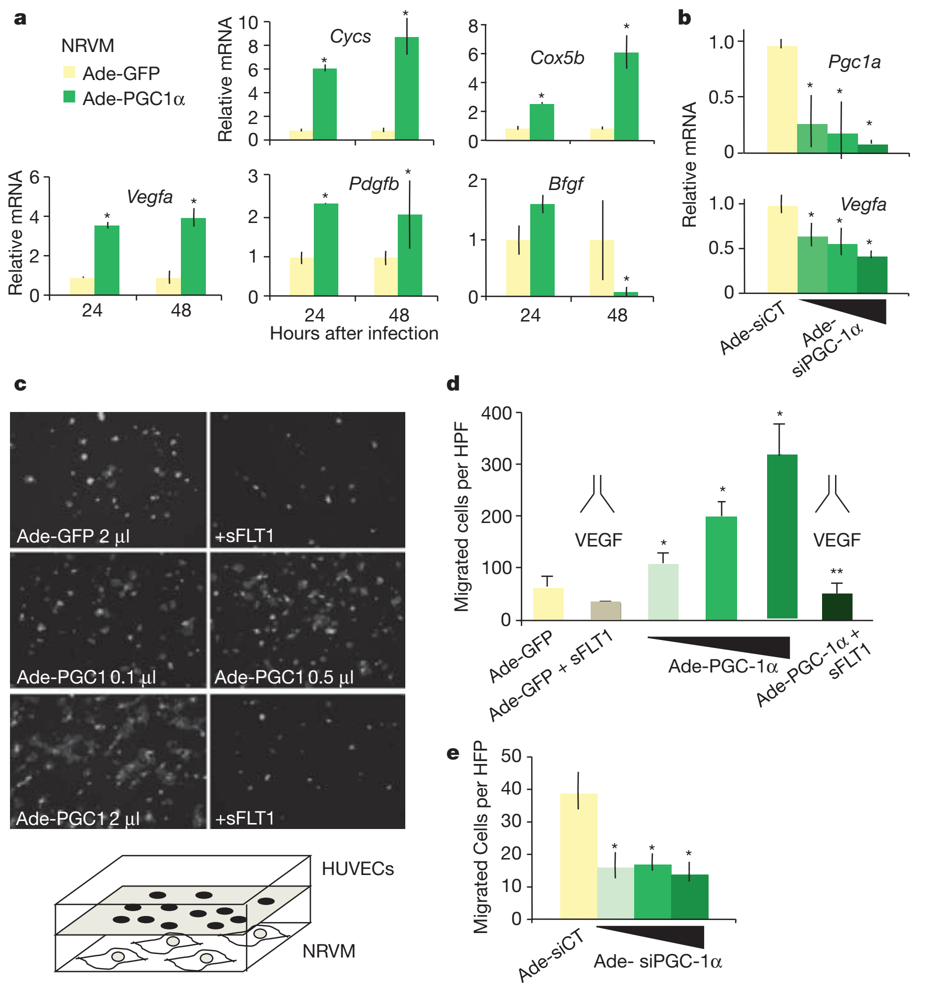
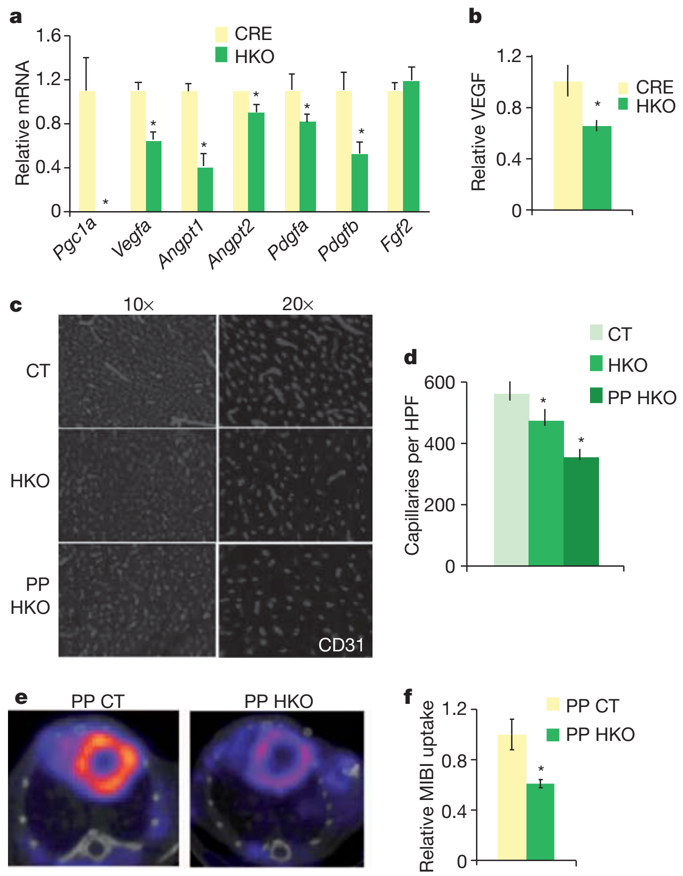


Figure 2 | PGC-1α regulates an angiogenic program in cardiomyocytes. a, b, Relative expression of mitochondrial genes (*Cycs* and *Cox5b*) and angiogenic genes (*Vegfa*, *Pdgfb* and *Bfgf*) in NRVMs infected with adenovirus expressing PGC-1α (Ade-PGC-1α) versus adenovirus expressing green fluorescent protein (Ade-GFP) (**a**) or siPGC-1α versus siCT (**b**). **c,** PGC-1α expression in NRVMs induces the migration of adjacent endothelial cells, and the migration is blocked by sFLT1. Representative images show phalloidin-stained endothelial cells that have migrated towards the NRVMs. The experimental procedure using the Transwell migration chamber is shown in the bottom panel. **d,** Data from the procedure in **c** are quantified. HPF, high power field. **e,** Knockdown of PGC-1α inhibits the migration of endothelial cells. Error bars are ± s.e.m. *$P < 0.05$ versus control. **$P < 0.05$ versus cells not treated with sFLT1.



from the pituitary gland. PGC-1α is known to increase ROS scavenging[20]. In cardiomyocytes, overexpression of PGC-1α strongly induced *Mnsod* mRNA and protein expression (Supplementary Fig. 3a, b). Conversely, MNSOD was repressed in PGC-1α HKO hearts, and levels of ROS were increased (Supplementary Fig. 3c, d). Thus, PGC-1α and STAT3 both regulate MNSOD and ROS in cardiomyocytes, suggesting that the absence of PGC-1α in the heart may, like the absence of STAT3, lead to prolactin-mediated anti-angiogenic effects. Prolactin had no direct effects on PGC-1α or VEGF expression in cardiac cells, and prolactin levels did not differ in heart or serum between wild-type and HKO animals (Supplementary Fig. 4).

These observations suggest that PGC-1α regulates two separate pro-angiogenic pathways in the heart—a VEGF pathway and a prolactin pathway—and that aberration of both pathways in PGC-1α HKO mice leads to PPCM (Fig. 4a). To test this idea directly, both pathways were rescued simultaneously: breeding HKO mice were treated with daily subcutaneous injections of VEGF protein and with bromocriptine supplementation in the water (Fig. 4b). This double treatment resulted in complete rescue of the PPCM in HKO females

Figure 3 | Mice lacking cardiac PGC-1α have reduced microvascular density that is worsened by pregnancy. a, b, Relative mRNA expression of several angiogenic factors (**a**) and VEGF protein levels (**b**) in HKO versus CRE control hearts. **c, d,** Vascular density in hearts from nulliparous or post-partum CT and HKO mice. Representative images stained for CD31 (also known as PECAM) are shown (**c**) and quantified (**d**). **e, f,** Reduced cardiac MIBI uptake in PP HKO versus PP CT animals. Representative SPECT/CT images are shown (**e**), and quantified (**f**). $n \geq 5$ for all groups. Error bars are ± s.e.m. *$P < 0.05$ versus control.

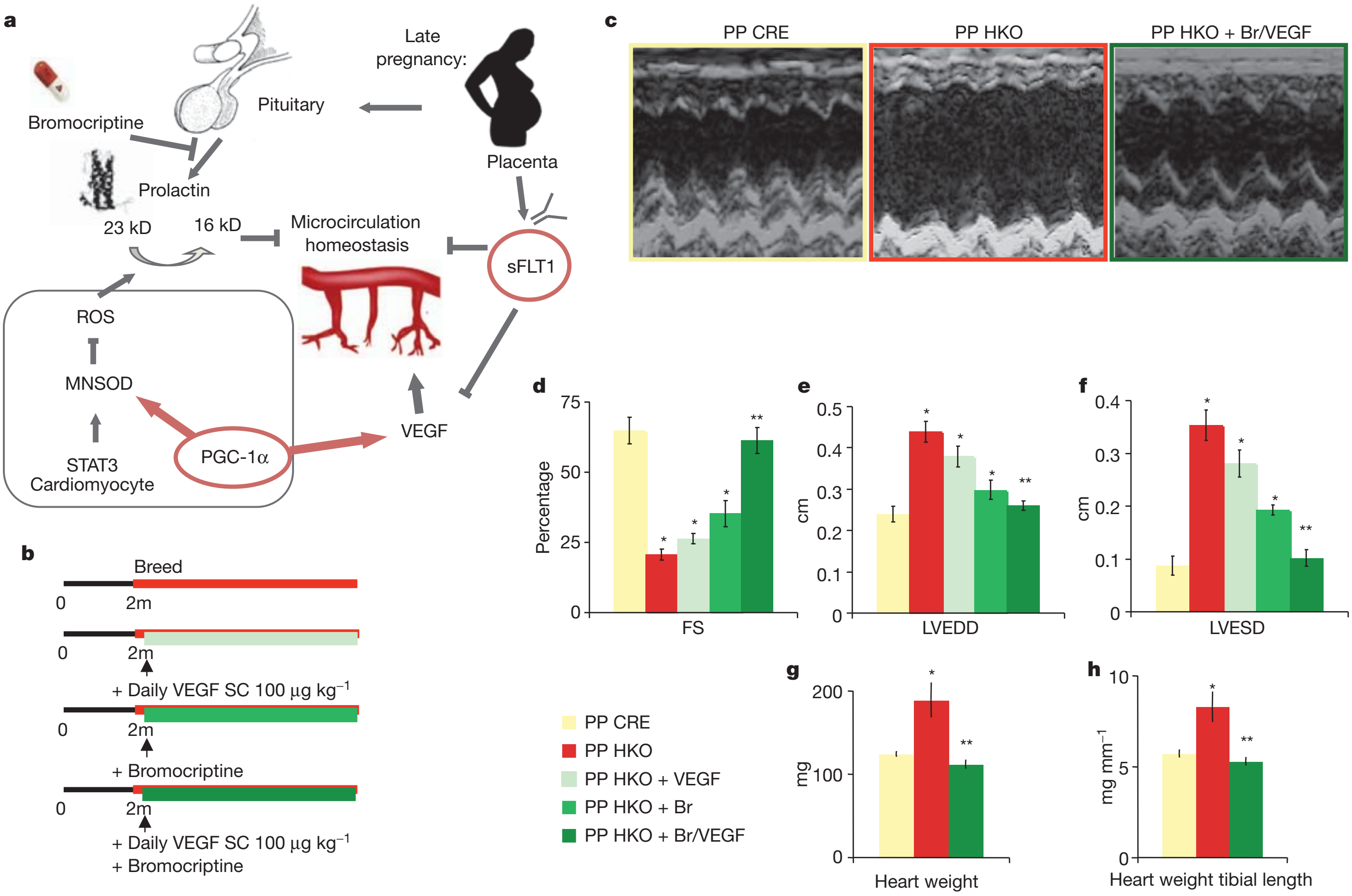


Figure 4 | Combined treatment with VEGF and bromocriptine rescues PPCM in PGC-1α HKO mice. **a**, Schema of the proposed role of cardiac PGC-1α in the regulation of cardiac angiogenesis, and in defending against pregnancy-induced anti-angiogenic factors. **b**, Experimental outline. m, months; SC, subcutaneous. **c**, Sample echocardiograms from PP HKO, PP CRE and PP HKO mice receiving both bromocriptine and VEGF treatments (PP HKO + Br/VEGF). **d–h**, Echocardiographic measures (**d–f**), heart weight (**g**) and heart weight:tibial length ratio (**h**), in PP mice of the indicated genotypes. $n \geq 5$ for all groups. Error bars are ± s.e.m. $*P < 0.05$ versus PP CRE control. $**P < 0.05$ versus PP HKO.

(Fig. 4c). After two pregnancies, heart weights and all echocardiographic indices of cardiac function (fractional shortening, LVEDD and LVESD) were normal in HKO mice (Fig. 4c–h). When only bromocriptine was used, some of the left ventricular dilation was prevented, but there was only minimal rescue of left ventricular function, showing that rescue of both pathways is necessary (Fig. 4c–h). Together, these data indicate that PPCM can be caused by an angiogenic imbalance and vascular dysfunction, at least in rodents.

Pre-eclampsia and cardiac dysfunction

To test the idea that anti-angiogenic signalling can cause cardiac dysfunction in pregnant women, we studied women with pre-eclampsia, in whom VEGF signalling is compromised owing to high serum levels of anti-angiogenic sFLT1[21] (see Supplementary Table 1 for patient characteristics). Cardiac function was evaluated non-invasively by measuring the myocardial performance index (MPI; also known as the Tei index) and other indices of cardiac function with cardiac echocardiography. MPI measures the relative duration of isovolaemic contraction and relaxation (Fig. 5a), and is a sensitive marker of diastolic function[22–24]. Women with pre-eclampsia had markedly increased serum levels of sFLT1 (Supplementary Fig. 5a, $P = 0.005$), as previously shown[21]. Notably, women with pre-eclampsia also had a markedly abnormal MPI (Fig. 5b and Supplementary Table 2, $P = 0.01$) and E/E' (Fig. 5c and Supplementary Table 2, $P = 0.02$), another sensitive measure of cardiac diastolic dysfunction that compares early diastolic mitral annulus (E') and transmitral (E) flow velocities[25]. Moreover, the MPI correlated with levels of circulating sFLT1 (Fig. 5d, $R = 0.59$,

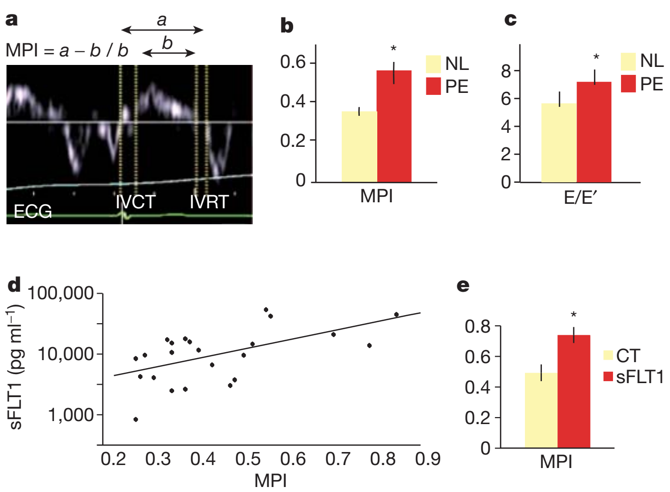


Figure 5 | Women with pre-eclampsia have depressed cardiac function that correlates with circulating sFLT1 levels, and sFLT1 causes cardiac dysfunction in mice. **a**, Sample tracing of echocardiographic tissue Doppler imaging. IVCT, isovolaemic contraction time; IVRT, isovolaemic relaxation time. **b**, **c**, Elevated MPI (**b**) and E/E' (**c**) in women with pre-eclampsia (PE) versus normal (NL) pregnancies. $P = 0.01$ and $P = 0.02$, respectively. **d**, Elevated MPI correlates with sFLT1 levels. $R = 0.59$. $P = 0.003$. **e**, Elevated MPI in pregnant mice infected with adenovirus expressing sFLT1. $P = 0.01$, $n \geq 5$ for all groups. $*P < 0.05$ versus control.

$P = 0.003$). Elevated blood pressure in the pre-eclamptic women (Supplementary Table 1) is unlikely to explain the worsening MPI, because MPI is thought to reflect cardiac function independently of blood pressure[22], and pregnant women with similar mild elevations of blood pressure but without pre-eclampsia have normal cardiac function[26]. Instead, these data suggest that elevated sFLT1 causes the diastolic dysfunction. To test this idea directly, sFLT1 was delivered systemically to pregnant mice by intravenous injection of adenoviruses expressing sFLT1, and MPI was examined using high-resolution murine echocardiography. sFLT1 caused significant increases in MPI in these mice within 10 days (Fig. 5e). These data, taken together with published observations in patients receiving anti-angiogenic therapies[10,11], strongly suggest that elevated sFLT1 causes cardiac dysfunction in women with pre-eclampsia. Although the left ventricular dysfunction recovers following delivery in many patients, a second insult in some women probably precipitates PPCM.

sFLT1 causes cardiomyopathy and is high in PPCM

The above observations strongly support the idea that PPCM can be induced by excess anti-angiogenic signalling, including the high expression of sFLT1 during late gestation seen both in women[27] and mice (Supplementary Fig. 5c, $P = 0.009$). To test this idea directly, sFLT1 was delivered systemically to nulliparous mice, as above. PGC-1α HKO mice that received sFLT1 developed profound cardiac failure within 3 weeks; these mice had increased cardiac weight and marked decreases in fractional shortening on echocardiography (Fig. 6a–c). This was accompanied by a marked drop in vascular density (Supplementary Fig. 6), although not in larger vessels (Supplementary Fig. 7). Wild-type mice also showed significant, though less extensive, decreases in vascular density and cardiac function after exposure to 3 weeks of sFLT1. Thus, sFLT1 alone is sufficient, even in the absence of pregnancy, to cause dramatic cardiomyopathy in the setting of a heart that is unable to withstand the anti-angiogenic insult.

To investigate further whether elevated sFLT1 levels in humans could be contributing to PPCM, plasma from women with PPCM was acquired 4–6 weeks post-partum and sFLT1 levels were measured. sFLT1 levels usually return to normal within 48–72 h after delivery[28]. sFLT1 levels were elevated in a large subset of these PPCM patients ($P = 0.002$), remaining up to five- or tenfold higher than the levels in control participants (Fig. 6d). Post-partum sFLT1 levels can remain slightly higher in subjects with pre-eclampsia[29,30], but the levels found here are notably higher. Thus, the findings are consistent with the idea that a substantial percentage of PPCM subjects have been exposed to pre-eclampsia, and that secretion of sFLT1 persists inappropriately post-partum. Indeed, in our own institution, 33% of the last 75 cases of PPCM were associated with pre-eclampsia (Fig. 6e, f), markedly more than the population rate of 3–5% (ref. 31). The persisting extra-placental source of sFLT1 in the post-partum period is not known, and may include placental remnants, circulating mononuclear cells[32] or shed syncytial microparticles[33].

Discussion

Our study shows that angiogenic imbalance in the heart during the peri-partum period may lead to PPCM in mice and in humans. The data indicate that PPCM is caused by a 'two-hit' combination of, first, systemic anti-angiogenic signals during late pregnancy and, second, a host susceptibility marked by insufficient local pro-angiogenic defences in the heart. The first hit explains why PPCM is a disease of the late gestational period, which is precisely when circulating anti-angiogenic factors such as sFLT1 peak in pregnancy[21,34]. Other pathways, such as prolactin or excess angiotensin II signalling, may also be involved[3,35]. The first hit is also worse in pre-eclampsia, which is characterized by markedly elevated sFLT1 levels. Associations between pre-eclampsia and PPCM have been well documented in many populations[1,2,36–41] (Supplementary Table 3). Interestingly, some studies involving women of African descent have not found an association between hypertensive disorders of pregnancy and PPCM[42], suggesting that there is ethnic variability in the pathogenesis of PPCM. It is also possible that PPCM with and without associated pre-eclampsia have different pathogeneses[43]. Overall, our data suggest that elevated sFLT1 levels in pre-eclampsia contribute to at least the PPCM that is associated with pre-eclampsia. We further propose that elevated sFLT1 levels in fact present a challenge to the myocardium in all pregnancies, thus explaining why the peri-partum period puts women at risk of developing heart failure, even in the absence of pre-eclampsia. Interestingly, other situations of elevated sFLT1 (twin pregnancies) and recurrent exposures to sFLT1 (multiple pregnancies) are also strong risk factors for PPCM even in the absence of pre-eclampsia[2,43,44].

Only a minority of women with pre-eclampsia develops PPCM, consistent with the existence of a second hit. Abnormal PGC-1α function is such an event in rodents, and it may also be a second hit in the case of humans. A number of previously identified processes may also constitute this second hit, including myocarditis, immune activation, viral infection and/or autoantibodies[43]. Interestingly, PGC-1α expression is repressed by inflammatory states in the heart and elsewhere[45,46], suggesting that many of the above processes that are implicated in PPCM may partly converge on PGC-1α. Consistent with this, we found repressed PGC-1α expression in cardiac samples from women with PPCM (Supplementary Fig. 8). Abnormal STAT3 function and ROS production[3] and genetic predispositions[47] may also be contributing factors.

In conclusion, the data presented here support the idea that PPCM is partly a two-hit vascular disease due to imbalances in angiogenic signalling, and that anti-angiogenic states such as pre-eclampsia or multiple gestation substantially worsen the severity of the disease. Our data may explain why pregnancy triggers PPCM, and also the long-standing epidemiological observation that pre-eclampsia is a risk factor for developing PPCM. Pro-angiogenic therapies such as exogenous VEGF121, or removal of sFLT1 itself[48], may therefore be beneficial in PPCM.

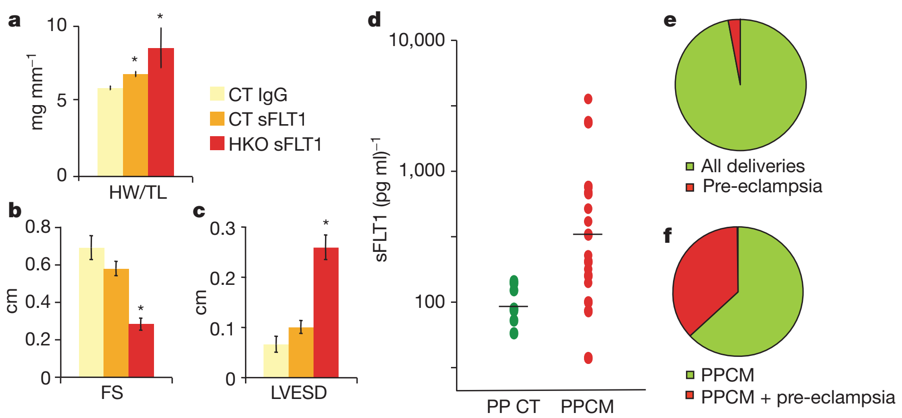


Figure 6 | sFLT1 is sufficient to induce cardiomyopathy in HKO mice, women with PPCM have elevated sFLT1 levels, and pre-eclampsia as a risk factor for PPCM. **a–c**, Heart weight:tibial length ratios (**a**), echocardiographic fractional shortening (**b**) and LVESD (**c**) in HKO mice injected with adenovirus expressing sFLT1, versus controls. **d**, Elevated sFLT1 levels in post-partum women with PPCM. $P = 0.002$. **e**, **f**, Prevalence of pre-eclampsia among all deliveries (**e**) and among women with PPCM (**f**) at Harvard teaching hospitals in the previous 9 years.

METHODS SUMMARY

All animal experiments were performed according to procedures approved by the Institutional Animal Care and Use Committee. Human soluble VEGF121 was a gift from Scios. Bromocriptine treatments were carried out as previously described[3]. Human studies were approved by the institutional review board of Beth Israel Deaconess Medical Center. Informed consent was obtained from all subjects. Angiogenic factor assays were performed with commercially available ELISA assays (R&D systems).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 September 2011; accepted 13 March 2012.
Published online 9 May 2012.

1. Pearson, G. D. *et al.* Peripartum cardiomyopathy: National Heart, Lung, and Blood Institute and Office of Rare Diseases (National Institutes of Health) workshop recommendations and review. *J. Am. Med. Assoc.* **283,** 1183–1188 (2000).
2. Sliwa, K., Fett, J. & Elkayam, U. Peripartum cardiomyopathy. *Lancet* **368,** 687–693 (2006).
3. Hilfiker-Kleiner, D. *et al.* A cathepsin D-cleaved 16 kDa form of prolactin mediates postpartum cardiomyopathy. *Cell* **128,** 589–600 (2007).
4. Rowe, G. C., Jiang, A. & Arany, Z. PGC-1 coactivators in cardiac development and disease. *Circ. Res.* **107,** 825–838 (2010).
5. Handschin, C. & Spiegelman, B. M. Peroxisome proliferator-activated receptor gamma coactivator 1 coactivators, energy homeostasis, and metabolism. *Endocr. Rev.* **27,** 728–735 (2006).
6. Arany, Z. *et al.* Transcriptional coactivator PGC-1α controls the energy state and contractile function of cardiac muscle. *Cell Metab.* **1,** 259–271 (2005).
7. Arany, Z. *et al.* Transverse aortic constriction leads to accelerated heart failure in mice lacking PPAR-γ coactivator 1α. *Proc. Natl Acad. Sci. USA* **103,** 10086–10091 (2006).
8. Arany, Z. *et al.* HIF-independent regulation of VEGF and angiogenesis by the transcriptional coactivator PGC-1α. *Nature* **451,** 1008–1012 (2008).
9. Chinsomboon, J. *et al.* The transcriptional coactivator PGC-1α mediates exercise-induced angiogenesis in skeletal muscle. *Proc. Natl Acad. Sci. USA* **106,** 21401–21406 (2009).
10. Kirk, R. Bevacizumab and heart failure. *Nature Rev. Clin. Oncol.* **8,** 124 (2011).
11. Uraizee, I., Cheng, S. & Moslehi, J. Reversible cardiomyopathy associated with sunitinib and sorafenib. *N. Engl. J. Med.* **365,** 1649–1650 (2011).
12. May, D. *et al.* Transgenic system for conditional induction and rescue of chronic myocardial hibernation provides insights into genomic programs of hibernation. *Proc. Natl Acad. Sci. USA* **105,** 282–287 (2008).
13. Carmeliet, P. *et al.* Impaired myocardial angiogenesis and ischemic cardiomyopathy in mice lacking the vascular endothelial growth factor isoforms VEGF164 and VEGF188. *Nature Med.* **5,** 495–502 (1999).
14. Bdolah, Y., Sukhatme, V. P. & Karumanchi, S. A. Angiogenic imbalance in the pathophysiology of preeclampsia: newer insights. *Semin. Nephrol.* **24,** 548–556 (2004).
15. Wu, Z. *et al.* Mechanisms controlling mitochondrial biogenesis and respiration through the thermogenic coactivator PGC-1. *Cell* **98,** 115–124 (1999).
16. Carmeliet, P. *et al.* Abnormal blood vessel development and lethality in embryos lacking a single VEGF allele. *Nature* **380,** 435–439 (1996).
17. Ferrara, N. *et al.* Heterozygous embryonic lethality induced by targeted inactivation of the VEGF gene. *Nature* **380,** 439–442 (1996).
18. Lehman, J. J. *et al.* Peroxisome proliferator-activated receptor γ coactivator-1 promotes cardiac mitochondrial biogenesis. *J. Clin. Invest.* **106,** 847–856 (2000).
19. Giordano, F. J. *et al.* A cardiac myocyte vascular endothelial growth factor paracrine pathway is required to maintain cardiac function. *Proc. Natl Acad. Sci. USA* **98,** 5780–5785 (2001).
20. St-Pierre, J. *et al.* Suppression of reactive oxygen species and neurodegeneration by the PGC-1 transcriptional coactivators. *Cell* **127,** 397–408 (2006).
21. Levine, R. J. *et al.* Circulating angiogenic factors and the risk of preeclampsia. *N. Engl. J. Med.* **350,** 672–683 (2004).
22. Bruch, C. *et al.* Tei-index in patients with mild-to-moderate congestive heart failure. *Eur. Heart J.* **21,** 1888–1895 (2000).
23. Tei, C. *et al.* New index of combined systolic and diastolic myocardial performance: a simple and reproducible measure of cardiac function—a study in normals and dilated cardiomyopathy. *J. Cardiol.* **26,** 357–366 (1995).
24. Poulsen, S. H., Jensen, S. E., Nielsen, J. C., Moller, J. E. & Egstrup, K. Serial changes and prognostic implications of a Doppler-derived index of combined left ventricular systolic and diastolic myocardial performance in acute myocardial infarction. *Am. J. Cardiol.* **85,** 19–25 (2000).
25. Kasner, M. *et al.* Utility of Doppler echocardiography and tissue Doppler imaging in the estimation of diastolic function in heart failure with normal ejection fraction: a comparative Doppler-conductance catheterization study. *Circulation* **116,** 637–647 (2007).
26. Melchiorre, K., Sutherland, G. R., Baltabaeva, A., Liberati, M. & Thilaganathan, B. Maternal cardiac dysfunction and remodeling in women with preeclampsia at term. *Hypertension* **57,** 85–93 (2010).
27. Venkatesha, S. *et al.* Soluble endoglin contributes to the pathogenesis of preeclampsia. *Nature Med.* **12,** 642–649 (2006).
28. Maynard, S. E. *et al.* Excess placental soluble fms-like tyrosine kinase 1 (sFlt1) may contribute to endothelial dysfunction, hypertension, and proteinuria in preeclampsia. *J. Clin. Invest.* **111,** 649–658 (2003).
29. Wolf, M. *et al.* Preeclampsia and future cardiovascular disease: potential role of altered angiogenesis and insulin resistance. *J. Clin. Endocrinol. Metab.* **89,** 6239–6243 (2004).
30. Saxena, A. R. *et al.* Increased sensitivity to angiotensin II is present postpartum in women with a history of hypertensive pregnancy. *Hypertension* **55,** 1239–1245 (2010).
31. Redman, C. W. & Sargent, I. L. Latest advances in understanding preeclampsia. *Science* **308,** 1592–1594 (2005).
32. Rajakumar, A. *et al.* Extra-placental expression of vascular endothelial growth factor receptor-1, (Flt-1) and soluble Flt-1 (sFlt-1), by peripheral blood mononuclear cells (PBMCs) in normotensive and preeclamptic pregnant women. *Placenta* **26,** 563–573 (2005).
33. Rajakumar, A. *et al.* Transcriptionally active syncytial aggregates in the maternal circulation may contribute to circulating soluble fms-like tyrosine kinase 1 in preeclampsia. *Hypertension* **59,** 256–264 (2012).
34. Noori, M., Donald, A. E., Angelakopoulou, A., Hingorani, A. D. & Williams, D. J. Prospective study of placental angiogenic factors and maternal vascular function before and after preeclampsia and gestational hypertension. *Circulation* **122,** 478–487 (2011).
35. Hubel, C. A. *et al.* Agonistic angiotensin II type 1 receptor autoantibodies in postpartum women with a history of preeclampsia. *Hypertension* **49,** 612–617 (2007).
36. Cruz, M. O., Briller, J. & Hibbard, J. U. Update on peripartum cardiomyopathy. *Obstet. Gynecol. Clin. North Am.* **37,** 283–303 (2010).
37. Elkayam, U. Clinical characteristics of peripartum cardiomyopathy in the United States: diagnosis, prognosis, and management. *J. Am. Coll. Cardiol.* **58,** 659–670 (2011).
38. Demakis, J. G. & Rahimtoola, S. H. Peripartum cardiomyopathy. *Circulation* **44,** 964–968 (1971).
39. Witlin, A. G., Mabie, W. C. & Sibai, B. M. Peripartum cardiomyopathy: an ominous diagnosis. *Am. J. Obstet. Gynecol.* **176,** 182–188 (1997).
40. Amos, A. M., Jaber, W. A. & Russell, S. D. Improved outcomes in peripartum cardiomyopathy with contemporary. *Am. Heart J.* **152,** 509–513 (2006).
41. Goland, S. *et al.* Evaluation of the clinical relevance of baseline left ventricular ejection fraction as a predictor of recovery or persistence of severe dysfunction in women in the United States with peripartum cardiomyopathy. *J. Card. Fail.* **17,** 426–430 (2011).
42. Fett, J. D., Christie, L. G., Carraway, R. D. & Murphy, J. G. Five-year prospective study of the incidence and prognosis of peripartum cardiomyopathy at a single institution. *Mayo Clin. Proc.* **80,** 1602–1606 (2005).
43. Ntusi, N. B. & Mayosi, B. M. Aetiology and risk factors of peripartum cardiomyopathy: a systematic review. *Int. J. Cardiol.* **131,** 168–179 (2009).
44. Bdolah, Y. *et al.* Twin pregnancy and the risk of preeclampsia: bigger placenta or relative ischemia? *Am. J. Obstet. Gynecol.* **198,** 428.e1–428.e6 (2008).
45. Schilling, J. *et al.* Toll-like receptor-mediated inflammatory signaling reprograms cardiac energy metabolism by repressing peroxisome proliferator-activated receptor γ coactivator-1 signaling. *Circ. Heart Fail.* **4,** 474–482 (2011).
46. Tran, M. *et al.* PGC-1α promotes recovery after acute kidney injury during systemic inflammation in mice. *J. Clin. Invest.* **121,** 4003–4014 (2011).
47. Horne, B.D., *et al.* Genome-wide significance and replication of the chromosome 12p11.22 locus near the PTHLH gene for peripartum cardiomyopathy. *Circ. Cardiovasc. Genet.* 4, 359–366..
48. Thadhani, R. *et al.* Pilot study of extracorporeal removal of soluble fms-like tyrosine kinase 1 in preeclampsia. *Circulation* **124,** 940–950 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements S.R. is supported by a Harvard Faculty Development and Diversity Award. S.S is supported by a John Hedley White grant. G.C.R. is supported by a Merck Fellowship. M.R.H. is supported by Harvard Catalyst, the Clinical and Translational Science Center, and Harvard University and affiliated academic health care centers. S.A.K. is an investigator of the Howard Hughes Medical Institute. Z.A. is supported by the NHLBI, the Smith Family Foundation, the Ellison Medical Foundation, the March of Dimes Foundation, and the Harvard Stem Cell Institute.

Author Contributions I.S.P. performed the majority of the mouse experimental work, with the assistance of G.C.R, L.L., N.K. and C.F. The clinical MPI study was preformed by S.R., S.S., J.S.R., M.R.H., J.M., F. M. and P.H. sFLT1 measurements were performed by S.R. MPI studies were performed by E.V.K. and S.D.B. The endothelial migration studies were performed by L.L. and C.J. Samples from women with PPCM were provided by J.B., F.d.M., I.T. and D.H.-K. These authors also provided input on the manuscript. The study was conceived and supervised by S.A.K. and Z.A. The experimental procedures were designed by Z.A., who also wrote the manuscript. All authors read and approved the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to Z.A. (zarany@bidmc.harvard.edu) or D.H.-K (hilfiker.denise@mh-hannover.de).

METHODS

Animal studies. All animal experiments were performed according to procedures approved by the Institutional Animal Care and Use Committee. Mice bearing floxed alleles of PGC-1α flanking exons 3 and 4, and mice containing the α-MHC::CRE transgene were gifts from B. Spiegelman[49] and M. Schneider[50], respectively. Mice were maintained on a standard rodent chow diet with 12-h light and dark cycles. For murine echocardiography, the chest hair was removed with a topical depilatory agent, and two-dimensional images were visualized using a Vivid FiVe echocardiography system (GE Medical Systems) on mice that were not anaesthetized. Parasternal short-axis projections were visualized and M-mode recordings at the mid-ventricular level were recorded. Heart rate, LVEDD and LVESD were measured in at least three beats from at least three recordings and averaged, and left ventricular fractional shortening was then calculated (fractional shortening = (LVEDD–LVESD)/LVEDD). For the high-resolution MPI studies, a VisualSonics 2100 echocardiography machine was used on mice anaesthetized with isoflurane, and the MPI was calculated using the manufacturer's software program. SPECT/CT imaging of mice was performed by the Longwood Area Small Animal Imaging Facility (SAIF). For the VEGF treatment studies, human VEGF121 (100 μg kg^{-1}) was injected subcutaneously daily, versus saline as the control. For the bromocriptine studies, bromocriptine was added to the drinking water. Mice were bred starting at the age of 8 weeks while receiving either bromocriptine in the drinking water or daily subcutaneous VEGF121, or both.

Cells and reagents. All reagents were from Sigma, unless otherwise indicated. Human soluble VEGF121 was a gift from Scios. Staining of capillaries was performed using anti-CD31 antibody (BD Pharmingen) or isolectin B4 (Vector Lab). Quantification of capillaries was performed computationally, using Volocity software (Improvision, PerkinElmer), on three random fields chosen from the septum of transverse sections from the mid-heart. Staining of arterioles was performed using anti-SMA antibody (Santa Cruz) and quantified similarly, using random low-power fields. All quantifications were performed blindly. Isolation and culture of primary NRVMs was performed as described. Cells were infected with adenovirus at a multiplicity of infection of 10× to 30×, and mRNA expression was measured 24 or 48 h later. The adenovirus expressing PGC-1α and sFLT1 have been described[51,52]. Prolactin, VEGF and sFLT1 ELISA assays were from R&D Systems. The thiobarbituic acid reactive substances (TBARS) assay was performed on cardiac extracts according to the manufacturer's instructions (Cayman).

Gene expression studies. Total RNAs were isolated from mouse tissue or cultured cells using the Trizol method (Invitrogen). Samples for real-time polymerase chain reaction (PCR) analyses were reverse transcribed (Invitrogen), and quantitative real-time PCR reactions were performed on the complementary DNAs in the presence of fluorescent dye (SYBR green) on a BioRad CFX 384 Touch real-time PCR detection system. DNA products of the expected size were confirmed for each primer pair.

Endothelial migration assay. NRVMs in 24-well plates were infected with adenovirus expressing GFP or PGC-1α for 34 h. bovine serum albumen (BSA) or sFLT1 (100 ng ml^{-1}) was added to the media for 12 h. Then, 5×10^4 cells of HUVECs at 5×10^4 were put on the upper compartment of transwells (8.0-μm pore size, Corning no. 3422) pre-warmed with EBM2 media overnight at 37 °C. HUVEC migration to the lower compartment of transwells was measured after 12 h. Migrated HUVECs were fixed with 4% paraformaldehyde in PBS for 20 min at 25 °C, cells remained on the upper compartment were removed with a cotton swab. Cells were blocked with 5% BSA in PBS 0.2% Tween (PBST) and stained with phallodin fluorescein isothiocyanate in PBST for 4 h to visualize filamentous actin. Transwell inserts were washed three times in PBST and mounted onto slides with 4′,6-diamidino-2-phenylindole (DAPI) mounting medium.

Human studies. The institutional review board of Beth Israel Deaconess Medical Center in Boston approved this study. Eligible women were enrolled after providing written informed consent from November 2009 to May 2010. Pregnant women at least 18 years of age with a singleton pregnancy of at least 24 weeks and less than 41 weeks, and either a diagnosis of pre-eclampsia or without any hypertensive disorder of pregnancy were eligible. Exclusion criteria included pre-existing cardiovascular disease, pulmonary disease and non-gestational diabetes mellitus. Participants were recruited after admission to labour and delivery, the ante-partum floor or during a routine prenatal visit. All clinical data were taken from medical records. The diagnosis of pre-eclampsia was based on the National High Blood Pressure Education Program Working Group definition, also endorsed by the American Congress of Obstetricians and Gynaecologists (ACOG). A maternal–fetal medicine specialist confirmed all diagnoses. Archived plasma samples from subjects with PPCM have been previously described[3]. Patients in both studies were predominantly Caucasian. Retrospective analyses of PPCM and pre-eclampsia in the Harvard teaching hospitals were performed using the Harvard Shared Health Research Information Network (SHRINE)[53], a de-identified repository of aggregate patient information.

Human echocardiography. Bedside transthoracic echocardiograms were performed using a Siemens X-300 (Mountainview) machine, by two expert echocardiographers using P5-1 Transducer. Images were obtained with the patient lying in the left lateral decubitus position and were reported according to the American Society of Echocardiography guidelines. Images were stored in a cine-loop format with three cardiac cycles of non-compressed data with electrocardiogram information. The echocardiographers performed a comprehensive examination, which included a complete two-dimensional and colour flow Doppler assessment of the left ventricle, right ventricle and intra-cardiac valves. Specifically: ejection fraction with visual quantitative estimation; trans-mitral pulse wave Doppler (E and A waves and deceleration time); Doppler tissue image (both medial and lateral mitral annulus were interrogated, and the final value of peak velocity of E' was calculated as the average of three velocities at each location); MPI, with the calculation performed off-line using a Siemens Syngo DICOM viewing station (Mountainview). The echocardiograms were de-identified before calculating MPI. Ejection fraction, MPI and E/E' ratios were calculated. Each image was analysed blind by one of two echocardiographers.

Angiogenic factor assays. Women consented to a blood draw at the time of the echocardiogram. All samples for the MPI study were collected in the ante-partum before the delivery, whereas samples in the PPCM study were collected 4–6 weeks post-partum. The samples were centrifuged at 1,900*g* for 8 min and plasma was collected and stored at −80° C. Samples were randomly ordered and analysed by a single person in a blind fashion. ELISA assays for sFLT1 were performed with commercially available kits (R&D systems). All assays were performed in duplicate and values were averaged. If >20% difference was observed between duplicate values, the samples were re-analysed.

Data and statistical analysis. SAS 9.2 (SAS institute) was used for data analysis. All tests were two sided, and P values of less than 0.05 were considered statistically significant. Data are presented as mean ± standard error, or median and inter-quartile ranges, as indicated. Comparisons were made using the two-tailed Student's *t*-test or the non-parametric Mann–Whitney test, as indicated.

49. Handschin, C. *et al.* Abnormal glucose homeostasis in skeletal muscle-specific PGC-1α knockout mice reveals skeletal muscle-pancreatic β cell crosstalk. *J. Clin. Invest.* **117,** 3463–3474 (2007).
50. Agah, R. *et al.* Gene recombination in postmitotic cells. Targeted expression of Cre recombinase provokes cardiac-restricted, site-specific rearrangement in adult ventricular muscle in vivo. *J. Clin. Invest.* **100,** 169–179 (1997).
51. Puigserver, P. *et al.* A cold-inducible coactivator of nuclear receptors linked to adaptive thermogenesis. *Cell* **92,** 829–839 (1998).
52. Kuo, C. J. *et al.* Comparative evaluation of the antitumor activity of antiangiogenic proteins delivered by gene transfer. *Proc. Natl Acad. Sci. USA* **98,** 4605–4610 (2001).
53. Weber, G. M. *et al.* The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J. Am. Med. Inform. Assoc.* **16,** 624–630 (2009).

ARTICLE

doi:10.1038/nature10939

Structure of the human κ-opioid receptor in complex with JDTic

Huixian Wu[1], Daniel Wacker[1], Mauro Mileni[1], Vsevolod Katritch[1], Gye Won Han[1], Eyal Vardy[2], Wei Liu[1], Aaron A. Thompson[1], Xi-Ping Huang[2], F. Ivy Carroll[3], S. Wayne Mascarella[3], Richard B. Westkaemper[4], Philip D. Mosier[4], Bryan L. Roth[2], Vadim Cherezov[1] & Raymond C. Stevens[1]

Opioid receptors mediate the actions of endogenous and exogenous opioids on many physiological processes, including the regulation of pain, respiratory drive, mood, and—in the case of κ-opioid receptor (κ-OR)—dysphoria and psychotomimesis. Here we report the crystal structure of the human κ-OR in complex with the selective antagonist JDTic, arranged in parallel dimers, at 2.9 Å resolution. The structure reveals important features of the ligand-binding pocket that contribute to the high affinity and subtype selectivity of JDTic for the human κ-OR. Modelling of other important κ-OR-selective ligands, including the morphinan-derived antagonists norbinaltorphimine and 5′-guanidinonaltrindole, and the diterpene agonist salvinorin A analogue RB-64, reveals both common and distinct features for binding these diverse chemotypes. Analysis of site-directed mutagenesis and ligand structure–activity relationships confirms the interactions observed in the crystal structure, thereby providing a molecular explanation for κ-OR subtype selectivity, and essential insights for the design of compounds with new pharmacological properties targeting the human κ-OR.

The four opioid receptors, μ, δ, κ and the nociceptin/orphanin FQ peptide receptor, belong to the class A (rhodopsin-like) γ subfamily of G-protein-coupled receptors (GPCRs)[1] with a common seven-transmembrane helical architecture, and are coupled predominantly to heterotrimeric G_i/G_o proteins. Activation of these receptors by endogenous or exogenous ligands is linked to a number of neuropsychiatric sequelae, including analgesia, sedation, depression, dysphoria and euphoria[2]. The three closely related subtypes, μ-OR, δ-OR and κ-OR, share ~70% sequence identity in their seven transmembrane helices (I–VII), with more variations in the extracellular loops (ECLs) and very little similarity in their amino and carboxy termini[2]. The majority of endogenous opioid peptides have a defined preference for specific subtypes, for example, endorphins act via δ-ORs and μ-ORs, whereas dynorphins preferentially activate κ-ORs. However, most exogenous and synthetic opioid ligands interact promiscuously (see the K_i Database; http://pdsp.med.unc.edu/pdsp.php), probably owing to the high degree of similarity among binding pockets of opioid receptors. Although decades of focused medicinal chemistry efforts have yielded reasonably selective ligands for all four ORs (see the K_i Database), there remains substantial interest in the development of subtype-selective agonists and antagonists.

Recent breakthroughs in elucidating high-resolution structures of GPCRs in complex with small-molecule[3–7] and peptide[8] ligands are providing details of their function[9], leading to numerous rational ligand discovery studies[10,11]. However, whereas most of these structures belong to the α subfamily of class A GPCRs[1], the highly diverse peptide-binding γ subfamily is represented only by the CXCR4 chemokine receptor[8]; additional structural coverage is needed to elucidate the repertoire of features[12] that define the pharmacological profile of this subfamily. The κ-OR, identified based on studies with the κ-type prototypic agonist ketocyclazocine[13], represents an attractive target for structure determination. Several κ-OR-selective partial agonists and antagonists have been developed as potential antidepressants, anxiolytics and anti-addiction medications[14], whereas a widely abused, naturally occurring hallucinogen—salvinorin A (SalA)—was also found to be a highly selective κ-OR agonist[15]. Although many κ-OR agonists and antagonists have not demonstrated desirable pharmacological properties, lacking specificity or displaying frank psychotomimetic actions in humans[14,16], some have been shown to be viable drug candidates. A κ-OR ligand in early stages of clinical development, JDTic (3*R*)-1,2,3,4-tetrahydro-7-hydroxy-N-[(1*S*)-1-[[(3*R*,4*R*)-4-(3-hydroxyphenyl)-3,4-dimethyl-1-piperidinyl]methyl]-2-methylpropyl]-3-isoquinolinecarboxamide), was originally designed as a novel selective κ-OR antagonist[17] that blocks the κ-OR agonist U50,488-induced antinociception, while not antagonizing μ-OR agonist-induced analgesia[18]. JDTic also displays robust activity in rodent models of depression, anxiety, stress-induced cocaine relapse, and nicotine withdrawal[18,19]. Here we report the crystal structure of a human κ-OR construct, κ-OR–T4 lysozyme (T4L), in complex with JDTic at 2.9 Å resolution. The results provide structural insights into the atomic details of molecular recognition and subtype selectivity of the κ-OR and related ORs, and should catalyse the structure-based design of advanced human κ-OR agonists and antagonists with improved pharmacological profiles and enhanced therapeutic efficacies.

Overall architecture of the κ-OR

Structural studies were carried out using an engineered human κ-OR construct (see Methods and Supplementary Fig. 1) and crystallized in cholesterol-doped monoolein lipidic cubic mesophase (see Methods). The construct used showed pharmacological behaviour similar to that of a native receptor expressed in HEK293T cells (Supplementary Tables 2 and 3). Data collection and refinement statistics are shown in Supplementary Table 1.

The structure of κ-OR–JDTic was determined at 2.9 Å in the $P2_12_12_1$ space group. The asymmetric unit consists of two receptors forming a parallel dimer (Fig. 1a). The dimer interface with ~1,100 $Å^2$

[1]Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA. [2]National Institute of Mental Health Psychoactive Drug Screening Program, Department of Pharmacology and Division of Chemical Biology and Medicinal Chemistry, University of North Carolina Chapel Hill Medical School, 4072 Genetic Medicine Building, Chapel Hill, North Carolina 27599, USA. [3]Center for Organic and Medicinal Chemistry, Research Triangle Institute, P.O. Box 12194, Research Triangle Park, North Carolina 27709, USA. [4]Department of Medicinal Chemistry, Virginia Commonwealth University, 800 E. Leigh Street, Richmond, Virginia 23298, USA.

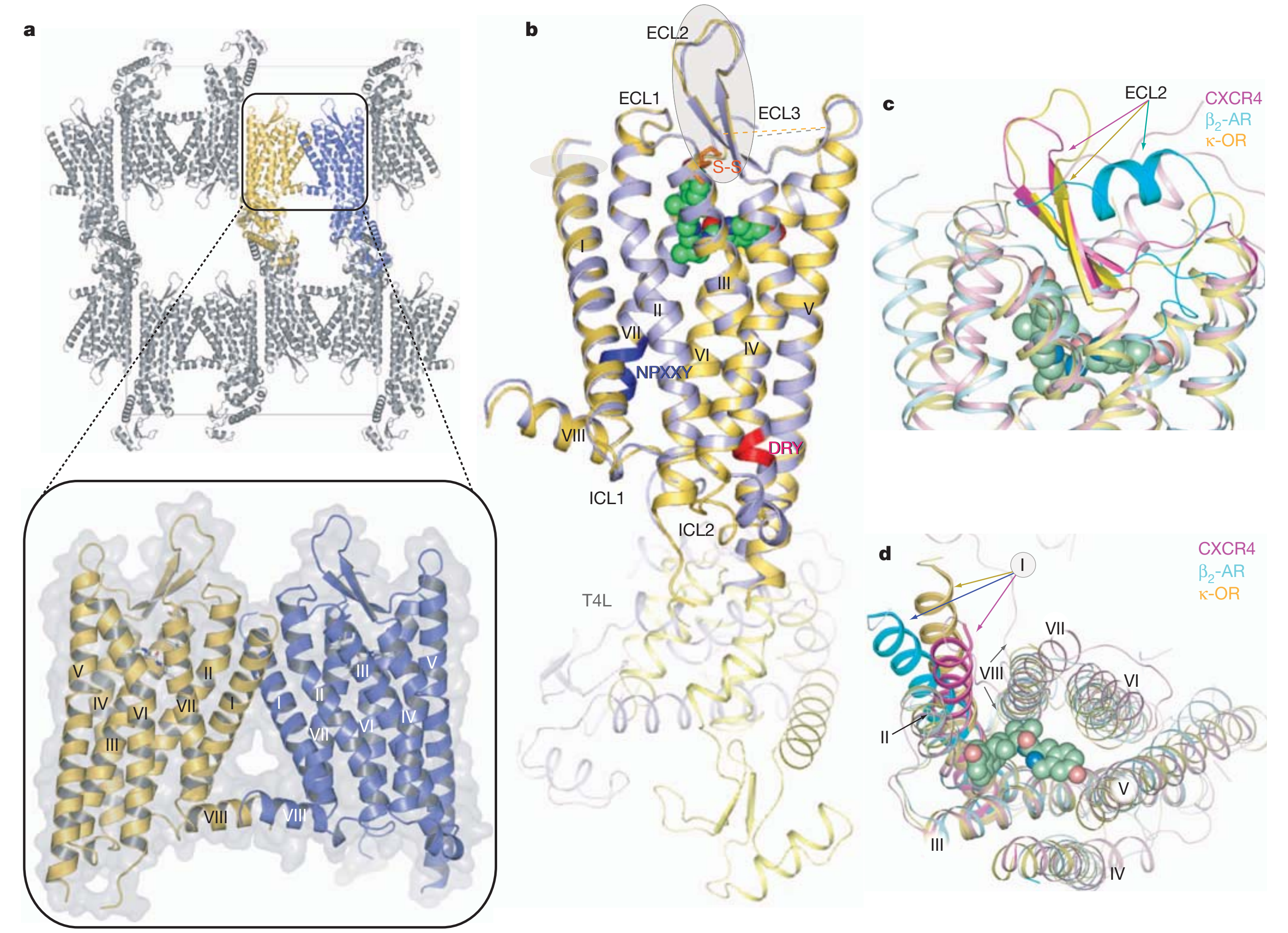


Figure 1 | Crystal packing and overview of the human κ-OR structure in complex with JDTic, and comparison with the inactive CXCR4 and β_2-AR structures. **a**, κ-OR–T4L crystal packing. The parallel dimer in one asymmetric unit is highlighted by the insert. **b**, Overall architecture of κ-OR–T4L in complex with JDTic. The A molecule (yellow) and B molecule (blue) from one asymmetric unit are aligned through the receptor part. The DRY and NPXXY motifs are highlighted in red and blue, respectively. JDTic is shown in a green sphere representation and the disulphide bond is coloured orange. **c**, **d**, Side (**c**) and extracellular (**d**) views of a structural alignment of the human κ-OR (yellow); CXCR4 (PDB accession 3ODU; magenta) and β_2-AR (PDB accession 2RH1; cyan). The graphics were created by PyMOL.

buried surface area is formed through contacts among helices I, II and VIII (Fig. 1a, insert). Previously, parallel receptor dimers have been identified in crystal structures of activated rhodopsin (involving helices I, II and VIII)[20], the β_2 adrenergic receptor (β_2-AR; cholesterol mediated)[3] and CXCR4 (involving helices IV, V and VI)[8]. Consistent with these crystallographic data, recent biochemical studies have suggested the existence of two dimerization interfaces: along helices IV and V (sensitive to receptor activation) and along helix I (insensitive to the state of activation)[21]. Although the orientations of the two T4L copies in the receptor monomers in one asymmetric unit differ by ~60° rotation, both copies of the receptor are highly similar (Fig. 1b) and will be treated identically except where otherwise noted.

The main fold of the human κ-OR consists of a canonical seven-transmembrane bundle of α-helices followed by an intracellular helix VIII that runs parallel to the membrane (Fig. 1a, b), resembling previously solved GPCR structures[3–8]. Structural comparison with other GPCRs suggests that human κ-OR has marked similarities in the ECL region with CXCR4, another peptide-binding receptor in the γ subfamily. In the seven-transmembrane region, however, the κ-OR structure is closer to aminergic receptors belonging to the α subfamily (alpha carbon root mean squared deviation (r.m.s.d.) ~2.3 Å for the β_2-AR, ~1.9 Å for the dopamine D3 receptor (D3R) and ~2.7 Å for CXCR4). The structure reveals distinctive features of the human κ-OR, including the following. First, conformation of the extracellular end of helix I deviates from the position observed in CXCR4, where the tip of helix I is pulled towards the transmembrane bundle by a disulphide bond between the N terminus and ECL3. Second, ECL2, the largest extracellular loop of the human κ-OR, forms a β-hairpin similar to that observed in CXCR4, despite the low sequence similarity in this domain between the two receptors. Conservation of this feature between these peptide receptors suggests that the β-hairpin could be a common motif in the ECL2 of other γ subfamily receptors, where interactions between ECL2 and their endogenous peptide ligands are deemed important for ligand recognition and selectivity[22]. Third, unlike other solved non-rhodopsin class A GPCRs that have more than one disulphide bond, the human κ-OR has only one formed between Cys $131^{3.25}$ (superscripts indicate residue numbering using the Ballesteros–Weinstein nomenclature[23]) and Cys 210, bridging ECL2 to the end of helix III. These two cysteines are conserved in all opioid receptors and this disulphide bond is the canonical one shared by most other solved class A GPCRs. Fourth, intracellular loop 2 (ICL2) adopts slightly different structures in the two κ-OR molecules in the asymmetric unit, involving a two-turn α-helix in molecule B, and only a one-turn α-helix in molecule A (Supplementary Fig. 2), possibly reflecting the conformational plasticity of this region[5]. Last, ECL3 of the κ-OR is disordered. Of the approximately 11 residues in this loop (residues 300–310), 6 residues in molecule A and 3 in molecule B do not have interpretable electron density.

A common feature of the class A GPCRs is the presence of a conserved sequence motif Asp/Glu$^{3.49}$-Arg$^{3.50}$-Tyr$^{3.51}$ (D/ERY) located at the cytoplasmic end of helix III. A salt bridge interaction between Arg$^{3.50}$ and Asp/Glu$^{6.30}$ from the cytoplasmic end of helix VI constitutes an 'ionic lock', which is thought to stabilize the inactive conformation of rhodopsin and other rhodopsin-like class A GPCRs[5,24], whereas its absence can enhance constitutive activity[6,23]. Although the human κ-OR lacks either of the acidic residues Asp/Glu at position 6.30, Arg 156$^{3.50}$ forms a hydrogen bond to another helix VI residue, Thr 273$^{6.34}$ (Supplementary Fig. 3a) in this inactive κ-OR structure, thereby conceivably stabilizing the inactive receptor conformation. The NPXXY motif located at the cytoplasmic side of helix VII, which is composed of Asn 326$^{7.49}$, Pro 327$^{7.50}$, Ile 328$^{7.51}$, Leu 329$^{7.52}$ and Tyr 330$^{7.53}$ in the κ-OR, is another highly conserved functional motif that is proposed to act as one of the molecular switches responsible for class A GPCR activation[25,26]. Comparison of the human κ-OR with inactive β_2-AR and A_{2A} adenosine receptor (A_{2A}AR) structures (Supplementary Fig. 3b) reveals a similar conformation of this motif in these receptors, thereby supporting the hypothesis that the observed κ-OR–JDTic complex structure corresponds to the inactive state. To establish further that JDTic stabilizes an inactive conformation, we evaluated its ability to modulate G_i/G_o-mediated and β-arrestin-mediated signalling in transfected HEK293T cells. We found that JDTic was devoid of agonist activity at both canonical and non-canonical pathways and completely blocked the effects of the prototypic agonist U69593 (Supplementary Fig. 4).

The κ-OR ligand-binding pocket

The κ-OR ligand-binding pocket displays a unique combination of key characteristics both shared with and distinct from those in the chemokine and aminergic receptor families. Although the human κ-OR binding pocket is comparatively large and partially capped by the ECL2 β-hairpin, as in CXCR4, it is also much narrower and deeper than in CXCR4 (Fig. 2c, d and Supplementary Fig. 5). In addition to a different set of side chains lining the pocket, the shape differences result from an approximately 4.5 Å inward shift of the extracellular tip of helix VI in the κ-OR as compared to CXCR4. The electron density clearly shows the position of the JDTic ligand (Supplementary Fig. 6), which reaches deep into the pocket to form ionic interactions with the Asp 138$^{3.32}$ side chain (Fig. 2a). The Asp$^{3.32}$ residue is conserved in all aminergic GPCRs, thereby having a critical role in the selectivity of aminergic receptors towards protonated amine-containing ligands. Likewise, Asp$^{3.32}$ is conserved in all opioid receptors, and modelling and mutagenesis studies[27] suggest that it has an essential role in anchoring positively charged κ-OR ligands.

Structural basis of JDTic selectivity

JDTic, developed as a derivative of the *trans*-(3*R*,4*R*)-4-(3-hydroxyphenyl)-3,4-dimethyl-1-piperidine scaffold[17], has exceptionally high affinity ($K_i = 0.32$ nM), potency ($K_i = 0.02$ nM in GTPγS assays)[17,28], long duration of action and a more than 1,000-fold selectivity for the human κ-OR as compared to other opioid receptor subtypes[28]. Extensive structure–activity relationship (SAR) analyses performed on JDTic analogues have yielded important insights into key determinants of JDTic activity[28–30], although reliable identification of the interaction mode(s) and contact residues of these ligands has not been feasible without a receptor crystal structure.

The crystal structure of κ-OR–JDTic shows a tight fit of the ligand in the bottom of the binding cleft (Fig. 2a), forming ionic, polar and extensive hydrophobic interactions with the receptor (Fig. 2b). The protonated amines in both piperidine and isoquinoline moieties of the ligand form salt bridges to the Asp 138$^{3.32}$ side chain (3.0 and 2.8 Å nitrogen–oxygen for molecule A, and 2.7 and 2.3 Å for molecule B, respectively). The piperidine amine is part of the original *trans*-(3*R*,4*R*)-dimethyl-4-(3-hydroxyphenyl)piperidine scaffold and is essential for opioid receptor antagonist activity[31]. SAR studies of JDTic analogues show that the isoquinoline nitrogen can be replaced by carbon, oxygen or sulphur atoms with only a ~10- to 50-fold

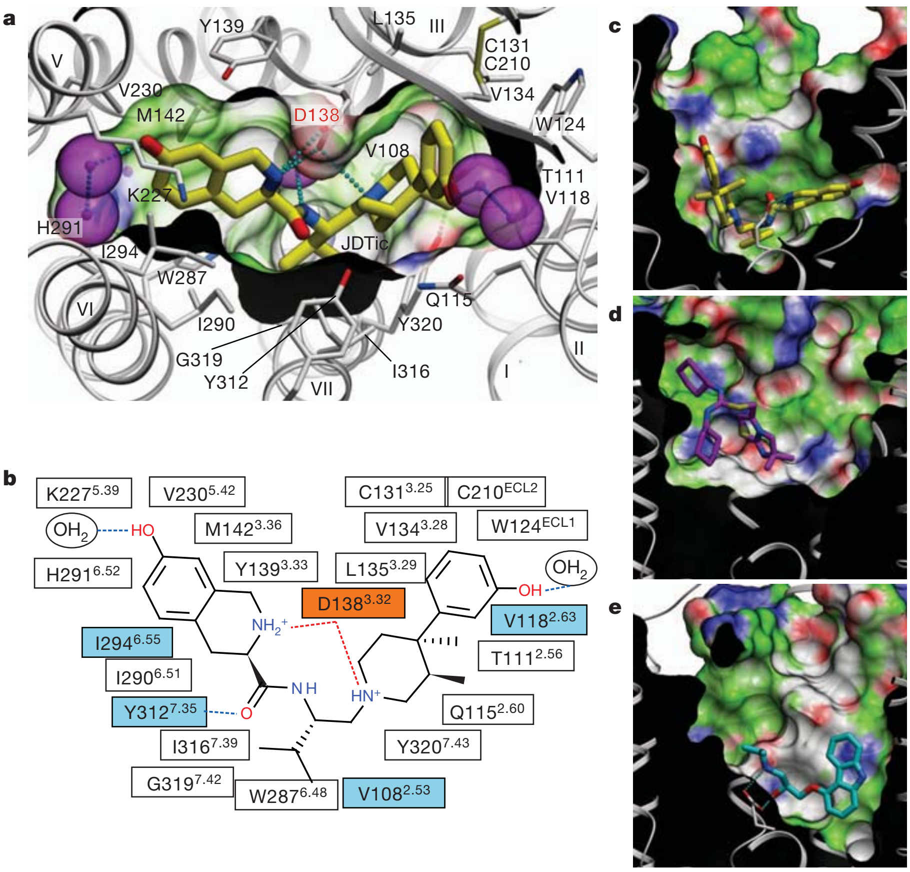


Figure 2 | Binding of the high-affinity selective antagonist JDTic in the human κ-OR crystal structure. **a**, Conformation of the binding pocket with JDTic shown by sticks with yellow carbons. The protein is displayed in cartoon representation looking down from the extracellular side, with the 22 contact residues within 4.5 Å from the ligand shown by white sticks. The pocket surface is shown as a semitransparent surface coloured according to binding properties (green: hydrophobic; blue: hydrogen-bond donor; red: hydrogen-bond acceptor). Salt bridges and hydrogen bonds are shown as dotted lines. Structured water molecules are shown as large magenta spheres. **b**, Diagram of ligand interactions in the binding pocket side chains at 4.5 Å cut-off. Salt bridges are shown in red and direct hydrogen bonds in blue dashed lines. Ballesteros–Weinstein numbering is shown as superscript. Residues that vary among the μ-OR, δ-OR and κ-OR subtypes are highlighted in cyan, and residue Asp 138$^{3.32}$ implicated in κ-OR-ligand binding by mutagenesis data, is highlighted orange. **c–e**, Side views of the sliced binding pocket in κ-OR–JDTic (**c**), CXCR4–IT1t (**d**) and β_2-AR–carazolol (**e**) complexes. The pocket surfaces are coloured as in panel **a**, the protein interior is black and the extracellular space is white. Ligands are shown as capped sticks with carbons coloured yellow (JDTic), magenta (IT1t) and cyan (carazolol). Asp$^{3.32}$ side chains in κ-OR–JDTic and β_2-AR–carazolol complexes are shown by thin sticks with grey carbons. The graphics were prepared using the ICM molecular modelling package (Molsoft LLC).

reduction in affinity[30]. Similar to the observed JDTic conformation in the κ-OR–JDTic complex, a V-shaped conformation was found in the small molecule X-ray crystal structure of JDTic, which showed its amino groups coordinating a water molecule (Supplementary Fig. 7a). Although several rotatable bonds within the JDTic molecule allow for the sampling of different conformations (see Supplementary Fig. 7b) and facilitate the ligand passage through the narrow binding pocket entrance, the anchoring-type interaction of two amino groups with Asp $138^{3.32}$ probably fixes the ligand in this characteristic V shape.

SAR studies have also underscored the importance of the distal hydroxyl groups on both the piperidine and isoquinoline moieties of JDTic, the removal of which did result in about a 100-fold reduction of affinity. A much smaller effect was observed upon methylation of these hydroxyls or their replacement by other polar groups[28]. These SAR results suggest the importance of water-mediated interactions between these two hydroxyl groups and the receptor. Indeed, although the crystal structure does not show direct hydrogen bonding with the receptor for both hydroxyl groups, there is clear electron density for several structured water molecules that mediate their polar interactions (Supplementary Fig. 6).

The κ-OR structure provides important clues for understanding the structural basis of the exceptional subtype selectivity of JDTic. Among many extensive contacts, JDTic interacts with four residues in the binding pocket that differ in other closely related opioid receptors, which are thought to contribute to the subtype selectivity of JDTic and other κ-OR-selective ligands[32] (human μ-OR and δ-OR amino acids are shown in parentheses, respectively): Val $108^{2.53}$ (Ala and Ala), Val $118^{2.63}$ (Asn and Lys), Ile $294^{6.55}$ (Val and Val) and Tyr $312^{7.35}$ (Trp and Leu) (Fig. 2b and Supplementary Fig. 8). Analysis of JDTic binding into κ-OR-based μ-OR and δ-OR homology models, as well as JDTic SAR results[17,28,30] (Supplementary Fig. 9), suggest that all described residues can contribute to the JDTic selectivity profile. Thus, changes in the Val $118^{2.63}$ side chain, where larger hydrophilic residues $\mathrm{Asn}^{2.63}$ and $\mathrm{Lys}^{2.63}$ are found in the human μ-OR and δ-OR, respectively, are likely to introduce unfavourable contacts with JDTic. Additionally, changing Tyr $312^{7.35}$ to the $\mathrm{Trp}^{7.35}$ and $\mathrm{Leu}^{7.35}$ residues found in the human μ-OR and δ-OR, respectively, is likely to result in the loss of an important polar interaction with the JDTic amide. The remaining two hydrophobic side-chain replacements, Val to Ala at position 2.53 and Ile to Val at position 6.55, may cause a reduction of the hydrophobic contact between JDTic and the receptor.

The isopropyl group of JDTic reaches deep into the orthosteric pocket to form a hydrophobic interaction with a conserved Trp $287^{6.48}$ side chain, possibly having a critical role in the pharmacological properties of this ligand. $\mathrm{Trp}^{6.48}$ is thought to be a key part of the activation mechanism in many class A GPCRs, including rhodopsin[26] and the $A_{2A}AR$[25], and similar hydrophobic contacts have been implicated in blocking activation-related conformational changes in the dark state visual rhodopsin by 11-*cis* retinal, and by inverse agonists in the $A_{2A}AR$ and D3R.

Binding of κ-OR-selective morphinans

Prior mutagenesis and modelling studies suggested that many small-molecule opioid ligands can interact with the κ-OR, as well as with the μ-OR and δ-OR, by forming a salt bridge with the highly conserved $\mathrm{Asp}^{3.32}$ (refs 33, 34). This is consistent with our mutagenesis studies (Supplementary Table 3) and flexible docking[35] of a series of morphine analogues, including selective κ-OR antagonists norbinaltorphimine (nor-BNI) and 5′-guanidinonaltrindole (GNTI) (Fig. 3 and Supplementary Fig. 10). To assess the compatibility of these bulky and rigid ligands with the observed κ-OR protein backbone conformation, we performed global energy optimizations of nor-BNI and GNTI in the binding cavity of κ-OR, keeping side chains of the binding pocket fully flexible. Multiple independent runs consistently resulted in low energy conformations with essentially identical poses and receptor contacts

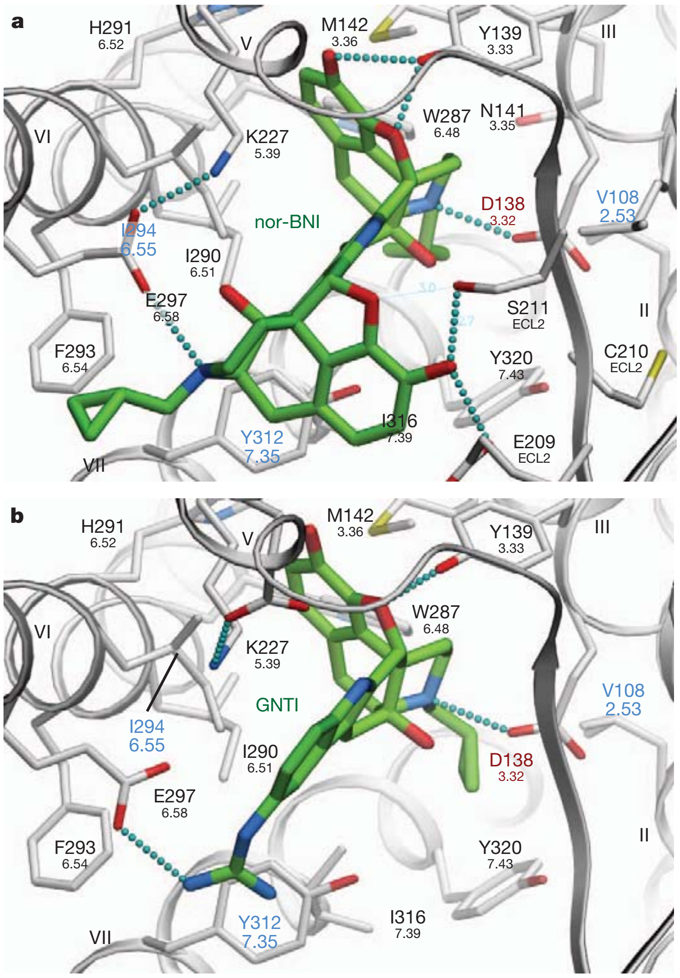


Figure 3 | Putative interaction modes of morphine-based high-affinity κ-OR-selective antagonists nor-BNI and GNTI. **a**, **b**, Interaction modes of nor-BNI (**a**) and GNTI (**b**). Ligands are depicted as capped sticks with green carbons, and contact side chains of the receptor within 4 Å from the ligand are shown with grey carbons. Key hydrogen bonds and salt bridges are indicated with small cyan spheres and residues unique to the κ-OR are labelled in blue. Residue Asp $138^{3.32}$, which also shows critical impact on GNTI and nor-BNI binding in mutagenesis studies, is highlighted in red. Ballesteros–Weinstein residue numbers are shown under the κ-OR residue numbers. The graphics were prepared using the ICM molecular modelling package (Molsoft LLC).

for the common naltrexone moieties of both nor-BNI and GNTI (r.m.s.d. 0.85 Å). In addition to a highly complementary van der Waals interface, both compounds formed an amino group salt bridge to the Asp $138^{3.32}$ side chain and a hydrogen bond to the Tyr $139^{3.33}$ side chain, both of which are important anchoring points for binding of morphine-based ligands, as supported by previous mutagenesis studies[34].

Moreover, unlike JDTic, both nor-BNI and GNTI compounds have a second basic moiety located more than 10 Å away from the first amino group (the second morphine moiety in nor-BNI and the guanidine moiety in GNTI). In the predicted models of κ-OR–nor-BNI/GNTI complexes, these additional amino groups of both ligands form a salt bridge with Glu $297^{6.58}$ located at the entrance to the ligand-binding pocket, which was previously characterized as a residue critical for subtype selectivity of κ-OR-selective morphinan derivatives[36]. This interaction is also supported by our mutagenesis results (Supplementary Table 3), where a Glu297Ala mutation induced a significant drop in both nor-BNI and GNTI binding affinity, but did not affect JDTic affinity. Hydrophobic interactions at the κ-OR-specific residue Ile 294 were also found for both nor-BNI and GNTI; consistent with our mutagenesis results (Supplementary

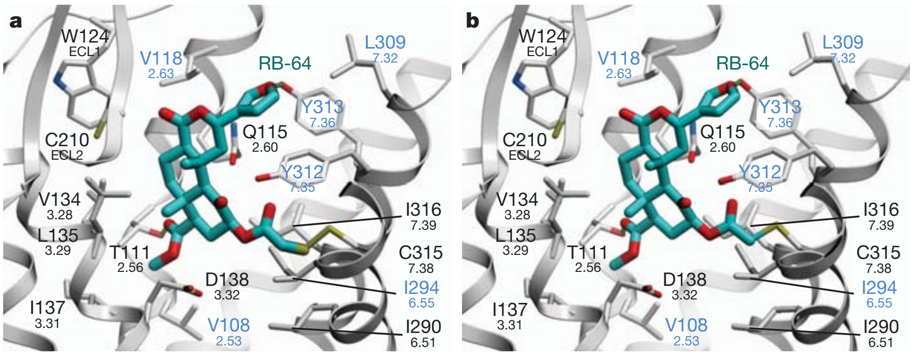


Figure 4 | Model of covalently bound RB-64. **a**, **b**, Putative binding mode of the RB-64 +463 AMU (**a**) and the RB-64 +431 AMU (**b**) adduct. Residues within 4 Å of the ligand are shown. Ligand, capped sticks/cyan carbons; κ-OR side chains, capped sticks; hydrogen bonds, small green spheres; κ-OR-unique residues are labelled in blue. Ballesteros–Weinstein residue numbers are shown under the κ-OR residue numbers. The graphics were prepared using the ICM molecular modelling package (Molsoft LLC).

Table 3) and suggesting that Ile 294 may also be important for developing human κ-OR-subtype-selective morphinan derivatives. Additional polar interactions with κ-OR-specific residues, Glu 209 and Ser 211 in ECL2, are found for nor-BNI, which may further enhance the κ-OR selectivity of this bulky ligand. Another side chain of the pocket, His $291^{6.52}$, which is involved in the highly conserved aromatic cluster around $\text{Trp}^{6.48}$ and thought to have a critical role in the receptor activation process[37], forms hydrophobic contacts with JDTic, nor-BNI and GNTI. His $291^{6.52}$ can be mutated to another aromatic residue, phenylalanine, without disrupting binding of these antagonists (Supplementary Table 3). The non-conservative His$291^{6.52}$Lys mutation, however, abolished binding of all tested ligands, probably because of the disruption of the aromatic cluster induced by the lysine side chain. Interestingly, the cyclopropyl moiety of both nor-BNI and GNTI in these binding poses has the same position as the isopropyl moiety of JDTic, making hydrophobic contact with the conserved residue Trp $287^{6.48}$. This cyclopropyl moiety is generally implicated in conversion of opioid agonists into antagonists (for example, agonist oxymorphone into antagonist naltrexone), and this effect may be partially explained by a direct interaction with the Trp $287^{6.48}$ side chain.

Overall, these structure-based docking results support the 'message–address' model[38] for morphine-based ligands nor-BNI and GNTI[36], which points to Glu $297^{6.58}$ as a key side chain that controls κ-OR selectivity by anchoring the 'address' moieties of these compounds. The crystal structure of the κ-OR–JDTic complex (Figs 2 and 3), however, demonstrates that even without an 'address' interaction with Glu $297^{6.58}$, more than a 1,000-fold subtype-selectivity to κ-OR can be achieved for JDTic and some of its derivatives. Importantly, then, the message–address hypothesis does not uniformly apply to all κ-OR-selective antagonists.

Binding of salvinorins

SalA, a naturally occurring diterpene from the widely abused hallucinogenic plant *Salvia divinorum*, represents an exceedingly potent (half-maximum effective concentration (EC_{50}) = 1 nM) and selective κ-OR agonist (>1,000-fold)[15]. SalA is unique compared to other κ-OR ligands in that it lacks a charged or polar nitrogen atom to anchor it in the binding pocket. Extensive site-directed mutagenesis, substituted cysteine-accessibility mutagenesis (SCAM) and SAR studies on SalA and its analogues have been performed, indicating (among others) that the 2-acetoxy moiety interacts with Cys $315^{7.38}$ (ref. 39). Possible modes of interaction between the cysteine-reactive and ultra-potent agonist and SalA analogue 22-thiocyanatosalvinorin A (RB-64; $K_i = 0.59$ nM; $EC_{50} = 0.077$ nM)[39], and the human κ-OR structure were thus evaluated. Exposure of κ-OR to RB-64 produces irreversibly bound, wash-resistant adducts that are tethered to Cys $315^{7.38}$ (ref. 39). As the thiocyanate group contains two electrophilic centres, two distinct adducts may be formed, increasing the mass by either 463 or 431 AMU. Docking studies using GOLD[40] predict that the salvinorin 2-position can access Cys $315^{7.38}$ while maintaining many of the interactions implicated by site-directed mutagenesis for SalA, providing a possible mechanism for the formation of the κ-OR–RB-64 adduct (Fig. 4, Supplementary Tables 4 and 5, and Supplementary Figs 11 and 12). Additionally, the docking results serve as a model of the initial recognition process of SalA-related agonists of the human κ-OR in an inactive state, although additional studies will be needed to fully elucidate the nature of the SalA-induced activation mechanism.

Conclusions

The κ-OR–JDTic crystal structure has uncovered a combination of key features shared with chemokine and aminergic GPCRs along with unique structural details characteristic of the opioid subfamily. The human κ-OR was crystallized as a parallel dimer with contacts involving helices I, II and VIII. Although the existence of GPCR dimers *in vivo* and their physiological relevance remain highly debatable, several distinct potential dimer interfaces are starting to emerge from crystallographic and biochemical studies. Such multiple dimerization interfaces may serve to support different functional pathways, as well as to promote oligomeric assembly of GPCRs. Analysis of ligand–receptor interactions has revealed important molecular details of the exceptionally high affinity and subtype selectivity of JDTic, a small-molecule antagonist with a broad therapeutic potential. The elucidation of a large binding cavity with a multitude of potential anchoring points begins to explain both the broad structural diversity of drugs targeting the human κ-OR and differences in their receptor interaction modes, as supported by differential effects of various site-directed mutations on the binding properties of chemically diverse prototypic ligands. The human κ-OR structure provides a long anticipated molecular framework for understanding opioid drug action, and thereby affords valuable new opportunities for the structure-based discovery of new drugs with ideal pharmacological properties.

METHODS SUMMARY

κ-OR–T4L was expressed in *Spodoptera frugiperda* (Sf9) cells. Ligand-binding and functional assays were performed as described in Methods. Receptor–ligand complexes were solubilized from washed Sf9 membranes using 1% (w/v) *n*-dodecyl-*β*-D-maltopyranoside (DDM) and 0.2% (w/v) cholesteryl hemisuccinate (CHS), and purified by immobilized metal ion affinity chromatography (IMAC), followed by reverse IMAC after cleaving N-terminal Flag–10×His tags by His-tagged tobacco etch virus (TEV) protease. The purified protein solution was mixed with monoolein and cholesterol in a ratio of 40%:54%:6% (w/w) to form lipidic cubic phase (LCP) from which the receptor was crystallized. Crystals were grown at 20 °C in 45 nl protein-laden LCP boluses overlaid by 800 nl of precipitant solutions as described in Methods. Crystals were harvested from the LCP matrix and flash frozen in liquid nitrogen. X-ray diffraction data were collected on the 23ID-B/D beamline (GM/CA CAT) at the Advanced Photon Source, Argonne, using a 10 μm minibeam at a wavelength of 1.0330 Å. Data collection, processing, structure solution and refinement are described in Methods. Modelling of JDTic analogues and κ-OR-selective

morphine derivatives nor-BNI and GNTI was performed using ICM-Pro; SYBYL-X 1.3 and GOLD Suite 5.1 were used to model RB-64 complexes, as described in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 31 December 2011; accepted 10 February 2012.
Published online 21 March 2012.

1. Fredriksson, R., Lagerstrom, M. C., Lundin, L. G. & Schioth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **63,** 1256–1272 (2003).
2. Waldhoer, M., Bartlett, S. E. & Whistler, J. L. Opioid receptors. *Annu. Rev. Biochem.* **73,** 953–990 (2004).
3. Cherezov, V. *et al.* High-resolution crystal structure of an engineered human β_2-adrenergic G protein-coupled receptor. *Science* **318,** 1258–1265 (2007).
4. Jaakola, V. P. *et al.* The 2.6 angstrom crystal structure of a human A_{2A} adenosine receptor bound to an antagonist. *Science* **322,** 1211–1217 (2008).
5. Chien, E. Y. *et al.* Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist. *Science* **330,** 1091–1095 (2010).
6. Warne, T. *et al.* Structure of a β_1-adrenergic G-protein-coupled receptor. *Nature* **454,** 486–491 (2008).
7. Shimamura, T. *et al.* Structure of the human histamine H_1 receptor complex with doxepin. *Nature* **475,** 65–70 (2011).
8. Wu, B. *et al.* Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* **330,** 1066–1071 (2010).
9. Rasmussen, S. G. *et al.* Crystal structure of the β_2 adrenergic receptor–Gs protein complex. *Nature* **477,** 549–555 (2011).
10. Katritch, V., Cherezov, V. & Stevens, R. C. Diversity and modularity of G protein-coupled receptor structures. *Trends Pharmacol. Sci.* **33,** 17–27 (2011).
11. Congreve, M., Langmead, C. J., Mason, J. S. & Marshall, F. H. Progress in structure based drug design for G protein-coupled receptors. *J. Med. Chem.* **54,** 4283–4311 (2011).
12. Kufareva, I., Rueda, M., Katritch, V., Stevens, R. C. & Abagyan, R. Status of GPCR modeling and docking as reflected by community-wide GPCR Dock 2010 assessment. *Structure* **19,** 1108–1126 (2011).
13. Martin, W. R., Eades, C. G., Thompson, J. A., Huppler, R. E. & Gilbert, P. E. The effects of morphine- and nalorphine- like drugs in the nondependent and morphine-dependent chronic spinal dog. *J. Pharmacol. Exp. Ther.* **197,** 517–532 (1976).
14. Carlezon, W. A. Jr, Beguin, C., Knoll, A. T. & Cohen, B. M. Kappa-opioid ligands in the study and treatment of mood disorders. *Pharmacol. Ther.* **123,** 334–343 (2009).
15. Roth, B. L. *et al.* Salvinorin A: a potent naturally occurring nonnitrogenous κ opioid selective agonist. *Proc. Natl Acad. Sci. USA* **99,** 11934–11939 (2002).
16. Walsh, S. L., Strain, E. C., Abreu, M. E. & Bigelow, G. E. Enadoline, a selective kappa opioid agonist: comparison with butorphanol and hydromorphone in humans. *Psychopharmacology* (*Berl.*) **157,** 151–162 (2001).
17. Thomas, J. B. *et al.* Identification of the first trans-(3*R*,4*R*)- dimethyl-4-(3-hydroxyphenyl)piperidine derivative to possess highly potent and selective opioid κ receptor antagonist activity. *J. Med. Chem.* **44,** 2687–2690 (2001).
18. Carroll, F. I. *et al.* Pharmacological properties of JDTic: a novel κ-opioid receptor antagonist. *Eur. J. Pharmacol.* **501,** 111–119 (2004).
19. Jackson, K. J., Carroll, F. I., Negus, S. S. & Damaj, M. I. Effect of the selective kappa-opioid receptor antagonist JDTic on nicotine antinociception, reward, and withdrawal in the mouse. *Psychopharmacology* (*Berl.*) **210,** 285–294 (2010).
20. Salom, D. *et al.* Crystal structure of a photoactivated deprotonated intermediate of rhodopsin. *Proc. Natl Acad. Sci. USA* **103,** 16123–16128 (2006).
21. Mancia, F., Assur, Z., Herman, A. G., Siegel, R. & Hendrickson, W. A. Ligand sensitivity in dimeric associations of the serotonin 5HT2c receptor. *EMBO Rep.* **9,** 363–369 (2008).
22. Wang, J. B., Johnson, P. S., Wu, J. M., Wang, W. F. & Uhl, G. R. Human κ opiate receptor second extracellular loop elevates dynorphin's affinity for human μ/κ chimeras. *J. Biol. Chem.* **269,** 25966–25969 (1994).
23. Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci.* **25,** 366–428 (1995).
24. Palczewski, K. *et al.* Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* **289,** 739–745 (2000).
25. Xu, F. *et al.* Structure of an agonist-bound human A_{2A} adenosine receptor. *Science* **332,** 322–327 (2011).
26. Standfuss, J. *et al.* The structural basis of agonist-induced activation in constitutively active rhodopsin. *Nature* **471,** 656–660 (2011).
27. Subramanian, G., Paterlini, M. G., Larson, D. L., Portoghese, P. S. & Ferguson, D. M. Conformational analysis and automated receptor docking of selective arylacetamide-based κ-opioid agonists. *J. Med. Chem.* **41,** 4777–4789 (1998).
28. Cai, T. B. *et al.* Synthesis and *in vitro* opioid receptor functional antagonism of analogues of the selective kappa opioid receptor antagonist (3*R*)-7-hydroxy-*N*-((1*S*)-1-{[(3*R*,4*R*)-4-(3-hydroxyphenyl)-3,4-dimethyl-1-pipe ridinyl]methyl}-2-methylpropyl)-1,2,3,4-tetrahydro-3-isoquinolinecarboxamide (JDTic). *J. Med. Chem.* **51,** 1849–1860 (2008).
29. Thomas, J. B. *et al.* Importance of phenolic address groups in opioid kappa receptor selective antagonists. *J. Med. Chem.* **47,** 1070–1073 (2004).
30. Runyon, S. P. *et al.* Analogues of (3*R*)-7-hydroxy-*N*-[(1*S*)-1-{[(3*R*,4*R*)-4-(3-hydroxyphenyl)-3,4-dimethyl-1-pipe ridinyl]methyl}-2-methylpropyl)-1,2,3,4-tetrahydro-3-isoquinolinecarboxamide (JDTic). Synthesis and *in vitro* and *in vivo* opioid receptor antagonist activity. *J. Med. Chem.* **53,** 5290–5301 (2010).
31. Zimmerman, D. M., Nickander, R., Horng, J. S. & Wong, D. T. New structural concepts for narcotic antagonists defined in a 4-phenylpiperidine series. *Nature* **275,** 332–334 (1978).
32. Vortherms, T. A., Mosier, P. D., Westkaemper, R. B. & Roth, B. L. Differential helical orientations among related G protein-coupled receptors provide a novel mechanism for selectivity. Studies with salvinorin A and the κ-opioid receptor. *J. Biol. Chem.* **282,** 3146–3156 (2007).
33. Surratt, C. K. *et al.* -mu opiate receptor. Charged transmembrane domain amino acids are critical for agonist recognition and intrinsic activity. *J. Biol. Chem.* **269,** 20548–20553 (1994).
34. Befort, K. *et al.* The conserved aspartate residue in the third putative transmembrane domain of the delta-opioid receptor is not the anionic counterpart for cationic opiate binding but is a constituent of the receptor binding site. *Mol. Pharmacol.* **49,** 216–223 (1996).
35. Totrov, M. & Abagyan, R. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins* **29,** 215–220 (1997).
36. Metzger, T. G., Paterlini, M. G., Portoghese, P. S. & Ferguson, D. M. Application of the message-address concept to the docking of naltrexone and selective naltrexone-derived opioid antagonists into opioid receptor models. *Neurochem. Res.* **21,** 1287–1294 (1996).
37. Chen, S. *et al.* Mutation of a single TMVI residue, Phe^{282}, in the β_2-adrenergic receptor results in structurally distinct activated receptor conformations. *Biochemistry* **41,** 6045–6053 (2002).
38. Chavkin, C. & Goldstein, A. Specific receptor for the opioid peptide dynorphin: structure–activity relationships. *Proc. Natl Acad. Sci. USA* **78,** 6543–6547 (1981).
39. Yan, F. *et al.* Structure-based design, synthesis, and biochemical and pharmacological characterization of novel salvinorin A analogues as active state probes of the κ-opioid receptor. *Biochemistry* **48,** 6898–6908 (2009).
40. Verdonk, M. L., Cole, J. C., Hartshorn, M., Murray, C. W. & Taylor, R. Improved protein-ligand docking using GOLD. *Proteins* **52,** 609–623 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by PSI:Biology grant U54 GM094618 (V.K., V.C., R.C.S.) for biological studies and structure production, NIH Roadmap grant P50 GM073197 (V.C., R.C.S.) for technology development and R01 DA017624 (B.L.R., E.V., R.B.M., P.D.M.), R01 DA027170 (B.L.R.), the NIMH Psychoactive Drug Screening Program Contract (B.L.R., X.-P.H.), the Michael Hooker Distinguished Chair of Pharmacology (B.L.R.), and the NIH grant R01 DA009045 (F.I.C.). D.W. is supported by a Boehringer Ingelheim Fonds PhD Fellowship. The JDTic X-ray structure was determined by C. George at the Laboratory for the Structure of Matter, Naval Research Laboratory. We thank J. Velasquez for help on molecular biology; T. Trinh, K. Allin and M. Chu for help on baculovirus expression; V. Setola for help with functional activity assays; J. Evans for help acquiring compounds; the National Institute of Drug Abuse Drug Supply Program for supplying JDTic and other opioid ligands used in these studies; K. Kadyshevskaya for assistance with figure preparation; E. Abola for assistance with manuscript preparation; A. Walker for assistance with manuscript preparation; J. Smith, R. Fischetti and N. Sanishvili for assistance in the development and use of the minibeam and beamtime at GM/CA-CAT beamline 23-ID at the Advanced Photon Source, which is supported by National Cancer Institute grant Y1-CO-1020 and National Institute of General Medical Sciences grant Y1-GM-1104.

Author Contributions H.W. assisted with protein expression, optimized the constructs, purified and crystallized the receptor in LCP, optimized crystallization conditions, grew crystals for data collection, collected the data and processed diffraction data, and prepared the manuscript. D.W. assisted with protein expression, purified the receptor, performed the thermal stability assay and assisted with preparing the manuscript. M.M. assisted with protein expression, purified the receptor, tested the JDTic compound, and performed the thermal stability assay. V.K. performed nor-BNI/GNTI-receptor docking and prepared the manuscript. G.W.H. processed diffraction data, solved and refined the structure and assisted with preparing the manuscript. E.V. created the initial tagged human κ-OR constructs and E.V. and X.-P.H. performed the ligand-binding and site-directed mutagenesis studies. W.L. assisted with construct optimization and crystallization in LCP. A.A.T. refined the structure and assisted with preparing the manuscript. F.I.C. and S.W.M. provided JDTic crystal structure, performed conformational studies of JDTic, and assisted with preparing the manuscript. R.B.W. and P.D.M. performed RB-64-receptor docking and prepared the manuscript. V.C. assisted with the crystallization in LCP, processed diffraction data, refined the structure and prepared the manuscript. B.L.R. suggested the JDTic compound for structural studies, supervised the pharmacology and mutagenesis studies and prepared the manuscript. R.C.S. was responsible for the overall project strategy and management and led the manuscript preparation and writing.

Author Information The coordinates and structure factors have been deposited in the Protein Data Bank under accession code 4DJH. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.C.S. (stevens@scripps.edu).

METHODS

Protein engineering for structural studies. Human κ-OR was engineered for structural studies by fusing lysozyme from T4 phage (T4L) into ICL3 (Gly 261–Arg 263) and further modified by N/C-terminal truncations (ΔGlu2–Ala42, ΔArg359–Val380) and a single point mutation Ile135$^{3.29}$Leu (see Supplementary Information). The resulting κ-OR–T4L construct was subsequently expressed in baculovirus-infected *Spodoptera frugiperda* (Sf9) insect cells.

Generation of κ-OR constructs for Sf9 expression. The human κ-OR (I135L) cDNA provided by the NIMH Psychoactive Drug Screening Program was cloned into a modified pFastBac1 vector (Invitrogen), designated as pFastBac1-833100, which contained an expression cassette with a haemagglutinin (HA) signal sequence followed by a Flag tag, a 10×His tag[3], and a TEV protease recognition site at the N terminus before the receptor sequence. Subcloning into the pFastBac1-833100 was achieved using PCR with primer pairs encoding restriction sites BamHI at the 5′ and HindIII at the 3′ termini of κ-OR wild type with subsequent ligation into the corresponding restriction sites found in the vector.

The κ-OR–T4L gene, based on the human κ-OR (I135L) and cysteine-free lysozyme from bacteriophage T4 (T4L C54T, C97A) sequences[41], included the following additional features: (1) residue Ser 262 at ICL3 of κ-OR was deleted by using standard QuickChange PCR; (2) Asn 2–Tyr 161 of T4L were inserted between Gly 261 and Arg 263 within the ICL3 region; and (3) N-terminal residues 2–42 and C-terminal residues 359–380 of κ-OR were truncated.

Expression and purification of κ-OR constructs. High-titre recombinant baculovirus (>10^9 viral particles per ml) was obtained using the Bac-to-Bac Baculovirus Expression System (Invitrogen) as previously described[5,8]. 25 μM of the antagonist naltrexone (NTX) and 5% Protein Boost Additive (PBA) were added to the system during expression. Cell suspensions were incubated for 4 days while shaking at 27 °C. Production of high-titre baculovirus stocks was performed as described before[5,8]. Sf9 cells at a cell density of 2–3 × 10^6 cells ml^{-1} were infected with P2 virus at a m.o.i. (multiplicity of infection) of 2. Cells were harvested by centrifugation at 48 h post-infection and stored at −80 °C until use.

Insect cell membranes were disrupted by thawing frozen cell pellets in a hypotonic buffer containing 10 mM HEPES, pH 7.5, 10 mM $MgCl_2$, 20 mM KCl and EDTA-free complete protease inhibitor cocktail tablets (Roche). Extensive washing of the raw membranes was performed by repeated centrifugation in the same hypotonic buffer (two to three times), and then in a high osmotic buffer containing 1.0 M NaCl, 10 mM HEPES, pH 7.5, 10 mM $MgCl_2$, 20 mM KCl and EDTA-free complete protease inhibitor cocktail tablets (three to four times), thereby separating soluble and membrane associated proteins from integral transmembrane proteins.

Washed membranes were resuspended into buffer containing 40 μM NTX, 2 mg ml^{-1} iodoacetamide, 150 mM NaCl and EDTA-free complete protease inhibitor cocktail tablets, and incubated at 4 °C for 1 h before solubilization. The membranes were then solubilized in 50 mM HEPES, pH 7.5, 150 mM NaCl, 1% (w/v) *n*-dodecyl-*β*-D-maltopyranoside (DDM, Anatrace), 0.2% (w/v) cholesteryl hemisuccinate (CHS, Sigma) and 20 μM NTX for 3 h at 4 °C. The supernatant was isolated by centrifugation at 160,000*g* for 40 min, and incubated in 30 mM buffered imidazole (pH 7.5), 1 M NaCl with TALON IMAC resin (Clontech) overnight at 4 °C. After binding, the resin was washed with 10 column volumes of Wash I Buffer (50 mM HEPES, pH 7.5, 800 mM NaCl, 10% (v/v) glycerol, 0.1% (w/v) DDM, 0.02% (w/v) CHS, 10 mM ATP, 10 mM $MgCl_2$ and 50 μM JDTic), followed by 6 column volumes of Wash II Buffer (50 mM HEPES, pH 7.5, 500 mM NaCl, 10% (v/v) glycerol, 0.05% (w/v) DDM, 0.01% (w/v) CHS, 50 mM imidazole and 50 μM JDTic). The protein was then eluted by 3 column volumes of Elution Buffer (50 mM HEPES, pH 7.5, 300 mM NaCl, 10% (v/v) glycerol, 0.03% (w/v) DDM, 0.006% (w/v) CHS, 250 mM imidazole and 50 μM JDTic). PD MiniTrap G-25 column (GE healthcare) was used to remove imidazole. The protein was then treated overnight with His-tagged AcTEV protease (Invitrogen) to cleave the N-terminal His-tag and Flag-tag. AcTEV protease and cleaved N-terminal fragment were removed by TALON IMAC resin incubation at 4 °C for 2 h for binding. The tag-less protein was collected as the TALON IMAC column flow-through. The protein was then concentrated to 40 mg ml^{-1} with a 100 kDa molecular weight cut-off Vivaspin centrifuge concentrator (GE healthcare). Protein purity and monodispersity were tested by SDS–PAGE and analytical size-exclusion chromatography (aSEC). Typically, the protein purity exceeded 95%, and the aSEC profile showed a single peak, indicative of receptor monodispersity.

Lipidic cubic phase crystallization. Protein samples of κ-OR in complex with JDTic were reconstituted into lipidic cubic phase (LCP) by mixing with molten lipid in a mechanical syringe mixer[42]. LCP crystallization trials were performed using an NT8-LCP crystallization robot (Formulatrix) as previously described[43]. 96-well glass sandwich plates (Marienfeld) were incubated and imaged at 20 °C using an automated incubator/imager (RockImager 1000, Formulatrix). Initial crystal hits were found from precipitant condition containing 100 mM sodium citrate pH 6.0, 30% (v/v) PEG400, 400 mM potassium nitrate. After extensive optimization, crystals of 30 μm × 10 μm × 5 μm to 60 μm × 20 μm × 10 μm size were obtained in 100 mM sodium citrate pH 5.8–6.4, 28–32% (v/v) PEG400, 350–450 mM potassium nitrate. Crystals were harvested directly from LCP matrix using MiTeGen micromounts and flash frozen in liquid nitrogen.

Data collection, structure solution and refinement. X-ray data were collected at the 23ID-B/D beamline (GM/CA CAT) at the Advanced Photon Source, Argonne, using a 10 μm minibeam at a wavelength of 1.0330 Å and a MarMosaic 300 CCD detector. Most crystals were invisible after flash freezing in liquid nitrogen, and a similar alignment and data-collection strategy was followed as previously described[44]. Among the several hundred crystal samples screened, most crystals diffracted to 2.8–3.5 Å resolution when exposed to 1–5 s of unattenuated beam using 1° oscillation. Data collection was limited to 5–10 frames per crystal, due to the fast onset of radiation damage in the microcrystals. Data were integrated, scaled and merged using HKL2000[45]. A 97% complete data set of κ-OR–T4L/JDTic (space group $P2_12_12_1$) at 2.9 Å resolution was obtained by merging data collected from 60 crystals. Initial phase information was obtained by molecular replacement with the program PHASER[46] using two independent search models of the polyalanine seven-transmembrane α-helices of CXCR4–IT1t (PDB accession 3ODU) and ensemble T4L models of β_2-AR–T4L (PDB accession 2RH1), A_{2A}AR–T4L (PDB accession 3EML), CXCR4–T4L (PDB accession 3ODU), D3R–T4L (PDB accession 3PBL) and H_1R–T4L (PDB accession 3RZE). Electron density refinement was performed with REFMAC5[47], autoBUSTER[48], and PHENIX[49] followed by manual examination and rebuilding of the refined coordinates in the program COOT[50] using both $|2F_o| - |F_c|$ and $|F_o| - |F_c|$ maps, as well as omit maps. The final model includes 287 residues of A chain (Ser 55–Gly 261, Arg 263–Ser 301, Ala 307–Pro 347) and 288 residues of B chain (Ser 55–Gly 261, Arg 263–Gly 300, Ser 305–Pro 347) of the κ-OR, and residues Asn 2–Tyr 161 of both A and B chains of T4L.

Ligand-binding assay. Membrane preparations, radioligand binding assays using ^{3}H-diprenorphine and data analyses were performed as previously described[39].

Modelling of high-affinity analogues of JDTic and morphine. Docking of high-affinity κ-OR-specific ligands was performed using an all-atom flexible receptor docking algorithm in ICM-Pro (MolSoft LLC) molecular modelling package as described previously[35]. Internal coordinate (torsion) movements were allowed in the side chains of the binding pocket, defined as residues within 10 Å distance of JDTic in the crystal structure. Other side chains and the backbone of the protein were kept as in the crystal structure. An initial conformation for each of the ligands was generated by Cartesian optimization of the ligand model in Merck Molecular Force Field. Docking was performed by placing the ligand in a random position within 5 Å from the entrance to the binding pocket and global conformational energy optimization of the complex[39,40]. To facilitate side-chain rotamer switches in flexible κ-OR models, the first 10^6 steps of the Monte Carlo (MC) procedure used 'soft' van der Waals potentials and high MC temperature, followed by another 10^6 steps with 'exact' van der Waals method and gradually decreasing temperature. A harmonic 'distance restraint' applied between an amino group of the ligand and carboxyl of Asp 138 side chain in the initial 10^6 steps was removed in the final 10^6 steps. At least 10 independent runs of the docking procedure were performed for each κ-OR ligand. The docking results were considered 'consistent' when at least 80% of the individual runs resulted in conformations clustered within a r.m.s.d. of <1 Å to the overall best energy pose of the ligand. All calculations were performed on a 12-core Linux workstation.

Modelling of RB-64. Modelling of RB-64 was performed using SYBYL-X 1.3 (Tripos) and GOLD 5.1 (Cambridge Crystallographic Data Centre)[40]. Default parameters were used except where noted. The structures of RB-64 and its κ-OR complexes were energy minimized using the Tripos Force Field (Gasteiger–Hückel charges, distance-dependent dielectric constant ε = 4, non-bonded interaction cut-off = 8 Å, energy gradient termination = 0.05 kcal/(mol × Å)). The κ-OR C315$^{7.38}$ χ_1 torsion angle was modified (χ_1 = +60.0°), orienting the sulfhydryl group towards the binding cavity. A docking distance constraint was used (C315 SG atom to thiocyanate sulphur atom distance, 2.0–6.0 Å; spring constant = 5.0). The Q115$^{2.60}$, D138$^{3.32}$, I290$^{6.51}$, I294$^{6.55}$, Y313$^{7.36}$ and I316$^{7.39}$ side chains were allowed to flex via rotamer library. The GoldScore fitness function was used with early termination disabled for 30 genetic algorithm runs. Poses were selected based on their GoldScore and ability to explain the relevant observed biochemical data. Stereochemical quality was assessed using PROCHECK.

41. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into β_2-adrenergic receptor function. *Science* **318,** 1266–1273 (2007).

42. Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nature Protocols* **4,** 706–731 (2009).
43. Cherezov, V., Peddi, A., Muthusubramaniam, L., Zheng, Y. F. & Caffrey, M. A robotic system for crystallizing membrane and soluble proteins in lipidic mesophases. *Acta Crystallogr. D* **60,** 1795–1807 (2004).
44. Cherezov, V. *et al.* Rastering strategy for screening and centring of microcrystal samples of human membrane proteins with a sub-10 μm size X-ray synchrotron beam. *J. R. Soc. Interface* **6** (Suppl. 5), S587–S597 (2009).
45. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
46. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Cryst.* **40,** 658–674 (2007).
47. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240–255 (1997).
48. Bricogne, G, *et al.* BUSTER v. 2.8.0 (Global Phasing, 2009).
49. Adams, P. D, *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
50. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).

ARTICLE

doi:10.1038/nature10954

Crystal structure of the μ-opioid receptor bound to a morphinan antagonist

Aashish Manglik[1], Andrew C. Kruse[1], Tong Sun Kobilka[1], Foon Sun Thian[1], Jesper M. Mathiesen[1], Roger K. Sunahara[2], Leonardo Pardo[3], William I. Weis[1,4], Brian K. Kobilka[1] & Sébastien Granier[1,5]

Opium is one of the world's oldest drugs, and its derivatives morphine and codeine are among the most used clinical drugs to relieve severe pain. These prototypical opioids produce analgesia as well as many undesirable side effects (sedation, apnoea and dependence) by binding to and activating the G-protein-coupled μ-opioid receptor (μ-OR) in the central nervous system. Here we describe the 2.8 Å crystal structure of the mouse μ-OR in complex with an irreversible morphinan antagonist. Compared to the buried binding pocket observed in most G-protein-coupled receptors published so far, the morphinan ligand binds deeply within a large solvent-exposed pocket. Of particular interest, the μ-OR crystallizes as a two-fold symmetrical dimer through a four-helix bundle motif formed by transmembrane segments 5 and 6. These high-resolution insights into opioid receptor structure will enable the application of structure-based approaches to develop better drugs for the management of pain and addiction.

Opium extracts from the plant *Papaver somniferum* have been used for therapeutic and recreational purposes for thousands of years. Opioid alkaloids and related pharmaceuticals are the most effective analgesics for the treatment of acute and chronic pain. They also represent one of the largest components of the illicit drug market worldwide, generating revenue of approximately $70 billion in 2009, much of which supports crime, wars and terrorism (UNODC World Drug Report 2011). Intravenous use of opioid drugs is a leading cause of death by overdose in Europe and North America, and a major contributing factor to the worldwide AIDS epidemic.

Morphine and codeine are the main active opioid alkaloids in opium. In humans, they act on the central nervous system to produce a wide range of effects including analgesia, euphoria, sedation, respiratory depression and cough suppression, and have peripheral effects such as constipation[1]. Gene disruption studies in mice show that the target for the majority of the effects of opioid alkaloids, whether beneficial or adverse, is the μ-OR[2]. The μ-OR belongs to the γ subfamily of class A G-protein-coupled receptors (GPCRs) with two closely related family members known as the δ- and κ-opioid receptors[3]. The μ-OR constitutes the main opioid target for the management of pain, acute pulmonary oedema, cough, diarrhoea and shivering[1]. However, opioid drugs are highly addictive, with the acetylated form of morphine, heroin, being the best-known example. Because of this, the clinical efficacy of opioid drugs is often limited by the development of tolerance and dependence.

Although both beneficial and adverse effects are attributable to activation of the μ-OR, they seem to be mediated by different downstream signalling and regulatory pathways. The μ-OR couples predominantly to Gi, the inhibitory G protein for adenylyl cyclase. μ-OR signalling through Gi is responsible for its analgesic properties[4]. After activation, the μ-OR undergoes phosphorylation and subsequently couples to arrestins, which have both regulatory and signalling functions[5]. Studies suggest that ligands with the greatest addictive potential, such as morphine, promote interactions with Gi more strongly than they promote interactions with arrestins[6]. These studies suggest that it may be possible to develop safer and more effective therapeutic agents targeting the μ-OR.

To understand better the structural basis for μ-OR function, we performed a crystallographic study of this receptor using the T4 lysozyme (T4L) fusion protein strategy developed previously[7] (Supplementary Fig. 1). Using the *in meso* crystallization method, we obtained crystals and collected diffraction data from 25 crystals of *Mus musculus* μ-OR–T4L protein bound to the irreversible morphinan antagonist β-funaltrexamine (β-FNA). The structure was solved by molecular replacement from a 2.8 Å data set.

Transmembrane architecture

The lattice for the μ-OR receptor shows alternating aqueous and lipidic layers with receptors arranged in parallel dimers tightly associated through transmembrane (TM) helices 5 and 6. More limited parallel interdimeric contacts through TM1, TM2 and helix 8 are observed between adjacent dimers (Supplementary Fig. 2).

As in other GPCRs, the structure of the μ-OR consists of seven TM α-helices that are connected by three extracellular loops (ECL1–3) and three intracellular loops (ICL1–3) (Fig. 1a). TM3 is connected to ECL2 by a conserved disulphide bridge between $C140^{3.25}$ (superscripts indicate Ballesteros–Weinstein numbers[8]) and C217. The morphinan ligand β-FNA (Fig. 1b, c) makes contacts with TM3, TM5, TM6 and TM7 (Fig. 1a), and the electron density observed in the structure confirms previous data identifying the $K233^{5.39}$ side chain as the site of covalent attachment[9] (Fig. 1c and Supplementary Fig. 3).

The intracellular face of the μ-OR closely resembles rhodopsin with respect to the relative positions of TM3, TM5 and TM6 (Supplementary Fig. 4). Nevertheless, like the β_2-adrenergic receptor (β_2-AR), there is no ionic bridge between the DRY sequence in TM3 and the cytoplasmic end of TM6. As with the β_2-AR, $R165^{3.50}$ forms a salt bridge with the adjacent $D164^{3.49}$ of the DRY sequence. $D164^{3.49}$ also engages in a polar interaction with R179 in ICL2, a feature that is similar to an interaction observed between $D130^{3.49}$ and S143 in ICL2

[1]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, California 94305, USA. [2]Department of Pharmacology, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. [3]Laboratori de Medicina Computacional, Unitat de Bioestadística, Universitat Autònoma de Barcelona, Barcelona 08193, Spain. [4]Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305, USA. [5]CNRS UMR 5203, INSERM U661, and Université Montpellier 1 et 2, Institut de Génomique Fonctionnelle, Montpellier 34094, France.

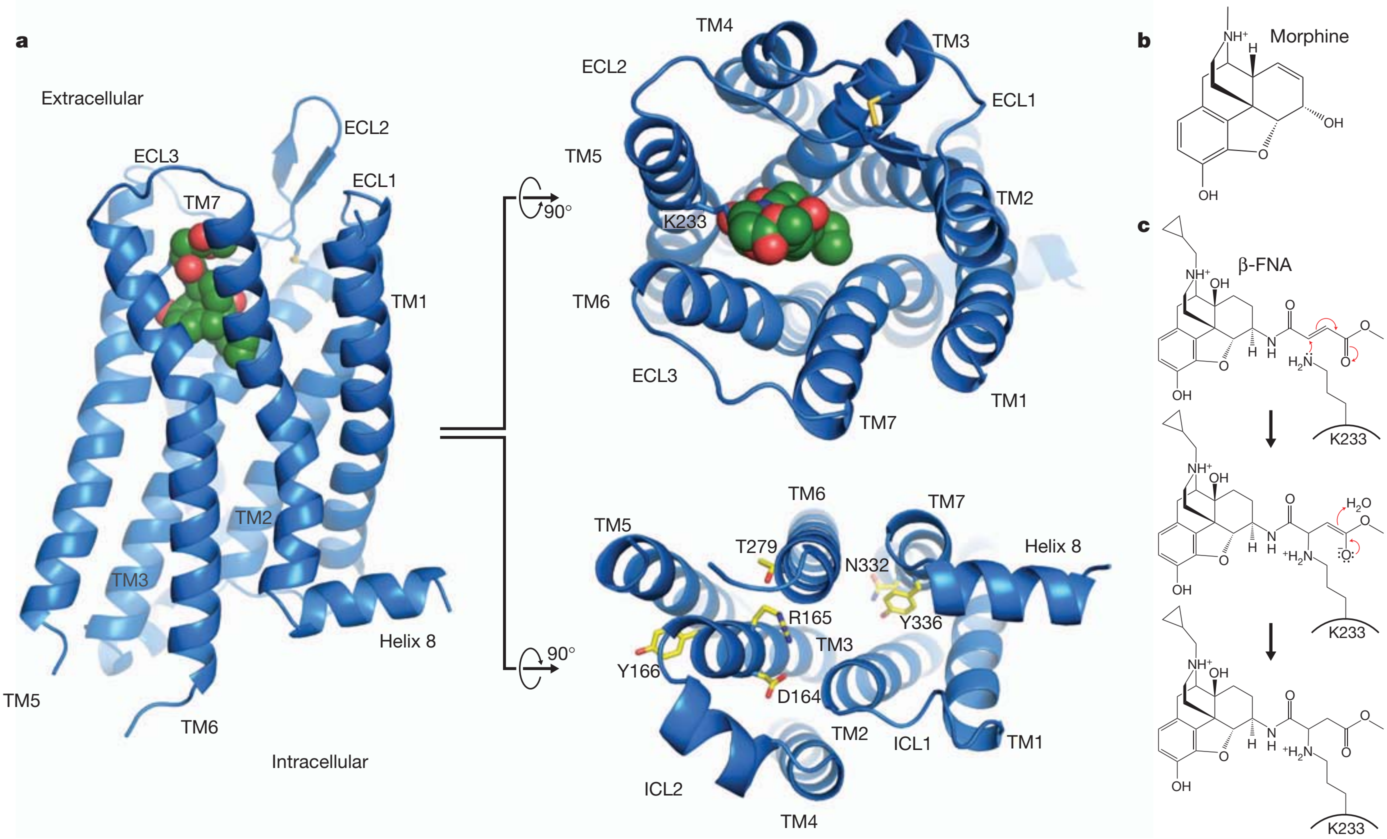


Figure 1 | Overall view of the μ-OR structure. **a**, Views from within the membrane plane (left), extracellular side (top) and intracellular side (bottom) show the typical seven-pass transmembrane GPCR architecture of the μ-OR. The ligand, β-FNA, is shown in green spheres. **b**, The chemical structure of morphine. **c**, The chemical structure of β-FNA and the chemical reaction with the side chain of K233[5.39] in the receptor are shown. β-FNA is a semisynthetic opioid antagonist derived from morphine, shown in **b**.

of the β_2-AR (Supplementary Fig. 4). In the μ-OR, it has been shown that the mutation of T279[6.34] to a lysine results in a constitutively active receptor[10]. This may be explained by a polar interaction observed in the crystal structure of the μ-OR between T279[6.34] and R165[3.50] (Supplementary Fig. 4). This interaction may stabilize the receptor in an inactive state.

An exposed ligand-binding pocket

In most available GPCR structures, the ligand is partially buried within the helical bundle by more superficial residues in TM segments and ECL2. The most extreme examples are the M2 and M3 muscarinic receptors[11,12], in which the ligand is covered with a layer of tyrosines (Fig. 2). This provides a structural basis for the very slow dissociation

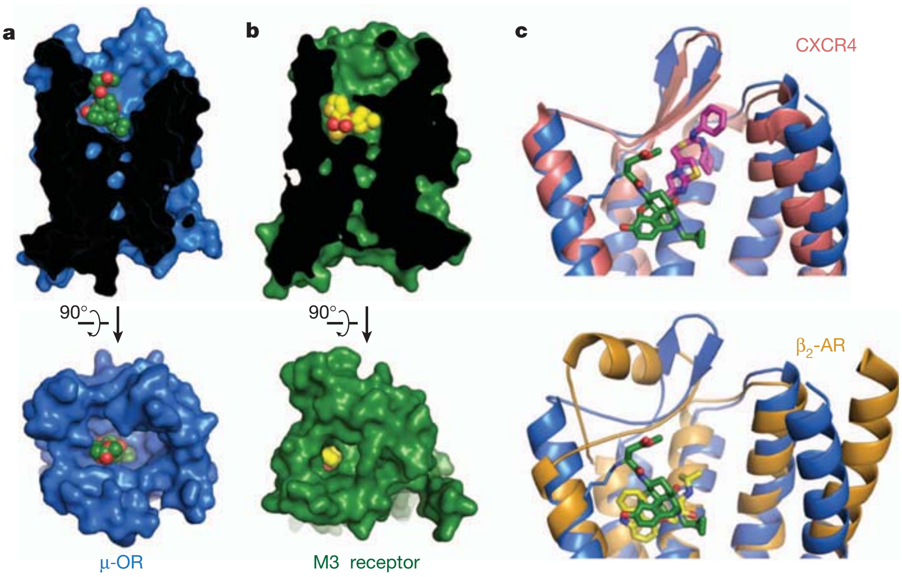


Figure 2 | Comparison of ligand-binding pockets. **a**, **b**, The binding pocket of the μ-OR (**a**) is wide and open above the ligand, in stark contrast to the deeply buried binding pocket of the muscarinic receptors, as exemplified by the M3 receptor (**b**). **c**, Top, the small-molecule antagonist IT1t (magenta) occupies a binding pocket closer to the extracellular surface of CXCR4 than β-FNA in μ-OR. Bottom, β-FNA is positioned more similarly to the distantly related aminergic receptors for the binding site of carazolol (yellow) in the β_2-AR.

kinetics of muscarinic antagonists. For example, the dissociation half-life of the clinically used drug tiotropium at the M3 receptor is 34.7 h and its dissociation constant (K_d) is 40 pM (ref. 13). By contrast, the binding pocket for β-FNA in the μ-OR is largely exposed to the extracellular surface (Fig. 2a). This may explain why extremely potent opioids such as buprenorphine, with an inhibition constant (K_i) of 740 pM, diprenorphine (K_i 72 pM), alvimopan (K_i 350 pM) and etorphine (K_i 230 pM) present rapid dissociation half-lives of 44 min, 36 min, 30 min[14] and less than 1 min (ref. 15), respectively. Therefore, although the affinity of high-affinity opioid ligands is comparable to tiotropium, the dissociation kinetics are considerably different. This feature of opioid ligands may explain why heroin overdoses are rapidly reversible by naloxone[16]. In addition, the extremely high potency and fast kinetics of etorphine agonism and diprenorphine antagonism allows for a system that is capable of rapid anaesthesia and prompt reversal in veterinary use. As a result, etorphine is a preferred anaesthetic (dose in the range of 5–20 μg kg^{-1}) for valuable racehorses and for captive and free-ranging mammals[17].

The μ-OR belongs to a subgroup of peptide GPCRs, and the closest published structure is that of the CXCR4 chemokine receptor[18] (root mean squared deviation (r.m.s.d.) value of 1.35 Å). In the μ-OR the morphinan ligand β-FNA binds much more deeply than the small-molecule CXCR4 antagonist IT1t and occupies a similar position as agonists and antagonists for the $β_2$-AR (r.m.s.d. value of 1.52 Å) and other monoamine receptors (Fig. 2c).

Binding pocket and opioid specificity

There are 14 residues within 4 Å of β-FNA. Nine of these have more direct interactions with the ligand (Fig. 3a–c), and are conserved in the κ-OR and δ-OR. $D147^{3.32}$ engages in a charge–charge interaction with the amine moiety of the ligand and hydrogen bonds with $Y326^{7.43}$ (both residues are strictly conserved in all the opioid receptor subtypes). Although $D147^{3.32}$ occupies the same position as the counterion in aminergic receptors, a sequence comparison shows that it is not conserved in other peptide receptors. $H297^{6.52}$ interacts with the aromatic ring of the morphinan group, but does not directly hydrogen bond with β-FNA as has been previously suggested[19]. However, the electron density suggests the presence of two water molecules that are well positioned to form a hydrogen-bonding network between $H297^{6.52}$ and the phenolic hydroxyl of the morphinan group (Fig. 3b, c).

A direct comparison with the δ-OR sequence also shows that of the 14 residues within 4 Å of the ligand, 11 are identical between μ-OR and δ-OR. The three differences are at μ-OR positions $E229^{ECL2}$, $K303^{6.58}$ and $W318^{7.35}$, which are Asp, Trp and Leu in the δ-OR, respectively. The substitution of leucine in δ-OR for $W318^{7.35}$ is highlighted in Fig. 3d. $W318^{7.35}$ was shown to be responsible for the binding selectivity of naltrindole, a δ-OR-selective antagonist and of [D-Pen2,D-Pen5]enkephalin (DPDPE), a δ-OR-selective peptide agonist[20]. In particular, the point mutation W318L markedly increases the affinity of both these ligands at the μ-OR. Positioning naltrindole (represented in Fig. 3d) into the μ-OR-binding pocket by

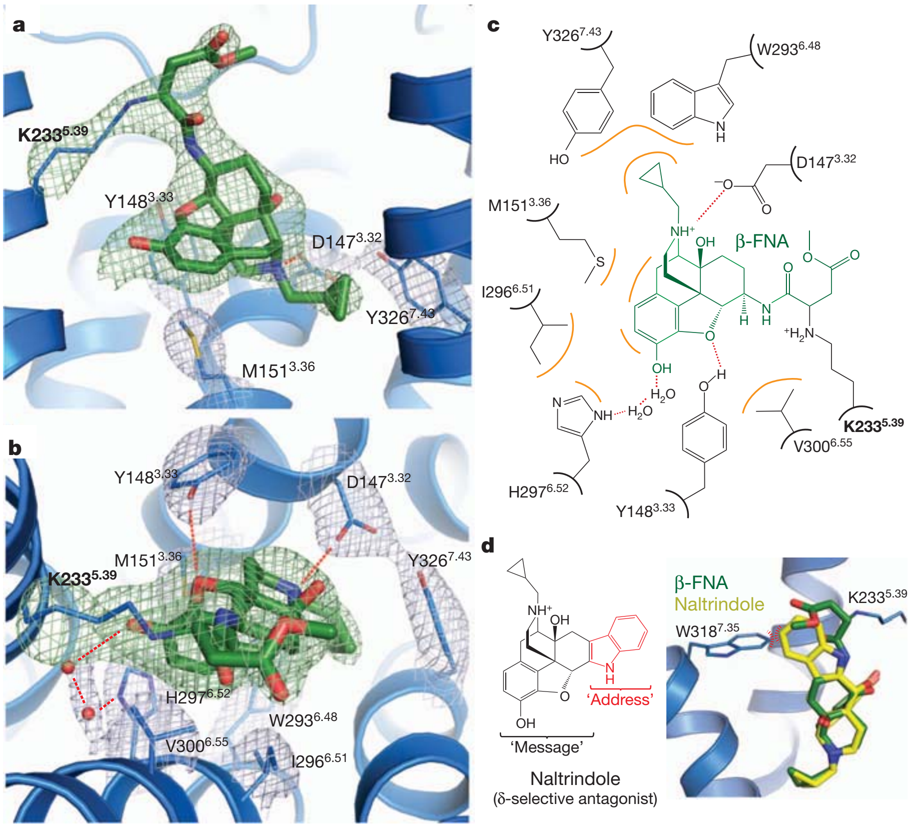


Figure 3 | Structural basis for morphinan ligand binding to the μ-OR. **a**, Side view of the ligand-binding pocket with polar interactions shown. TM6 is excluded from this view. The electron density used to position interacting side chains is shown in light blue coloured mesh depicting the $2F_o - F_c$ electron density contoured at 1.3σ. Green mesh depicts an omit map of β-FNA and $K233^{5.39}$ side-chain atoms contoured at 3.0σ. **b**, Binding pocket viewed from the extracellular surface. Water molecules are shown as red spheres, with the accompanying electron density shown in light blue mesh. **c**, The binding site is diagrammed, showing the chemical structure of β-FNA (green) covalently bound to the receptor through $K233^{5.39}$ (bold). Hydrophobic interactions are shown in orange and polar contacts with red dotted lines. $V300^{6.55}$ and $I296^{6.51}$ form extensive hydrophobic contacts with the back face of the ligand (not shown). Two water molecules are positioned between $H297^{6.52}$ and the phenolic group of β-FNA. **d**, The δ-OR-selective ligand naltrindole includes an indole group that would clash with $W318^{7.35}$ in μ-OR, but not with the leucine found in the equivalent position in δ-OR. The indole has been described as an 'address' to target the ligand to δ-OR, whereas its efficacy ('message') is determined by the morphinan group on the left[40].

superimposition of its morphinan group on that of β-FNA shows that naltrindole would clash with the W318 side chain in μ-OR (Fig. 3d), whereas the leucine in this position of δ-OR would probably accommodate naltrindole without requiring structural rearrangement.

Endomorphins 1 and 2 are small peptides isolated from brain that were shown to have the highest affinity (low nM range) and the highest selectivity profile for the μ-OR receptor[21]. For instance, endomorphin 1 exhibits 4,000- and 15,000-fold selectivity for μ-OR over δ-OR and κ-OR, respectively[21]. Although little is known about the determinants of endomorphin binding, mutagenesis studies suggest that the μ-OR-selective synthetic peptide agonist [D-Ala2,*N*-MePhe4,Gly-ol5] enkephalin (DAMGO) occupies a space that overlaps with the β-FNA-binding pocket but also extends beyond this site[22]. Sites of mutations that impair DAMGO binding include $H297^{6.52}$, positioned near the bottom of the β-FNA pocket, as well as $K303^{6.58}$, $W318^{7.35}$ and $H319^{7.36}$, positioned above the β-FNA-binding pocket (Supplementary Fig. 5). Given the residues involved in DAMGO binding to μ-OR, opioid peptides probably make both polar and non-polar contacts within the μ-OR-binding pocket. This feature of opioid peptide binding is also reflected in the lack of a highly charged surface within the μ-OR-binding pocket compared with that of the CXCR4 receptor[18].

Oligomeric arrangement of μ-OR

The structure of μ-OR shows receptor molecules intimately associated into pairs along the crystallographic two-fold axis through two different interfaces (Fig. 4a, b). The first interface is a more limited parallel association mediated by TM1, TM2 and helix 8, with a buried surface area of 615 Å^2 (Fig. 4d and Supplementary Fig. 6). The second and more prominent interface observed in the μ-OR crystal structure is comprised of TM5 and TM6 (Fig. 4c). In this case, within each μ-OR–μ-OR pair, the buried surface area for a single protomer is 1,492 Å^2. This represents 92% of the total buried surface between μ-OR–T4L molecules, indicating that the comparatively small 114 Å^2 buried surface contributed by T4L is unlikely to drive the contact (Supplementary Fig. 7). This suggests that the pairwise association of receptor monomers may represent a physiological opioid receptor dimer or higher-order oligomer, the existence of which is supported by previous biochemical, pharmacological and cell biological studies[23].

Recent computational and biochemical studies have indicated the potential role of TM4 and TM5 in the interaction between δ-OR receptors[24]. More generally, oligomers have been observed for a large number of GPCRs (recently reviewed in ref. 25). Some of these studies have shown that TM5 and TM6 peptides can disrupt dimers of the β_2-AR and V2 vasopressin receptor[26,27], and recent crosslinking experiments with the M3 muscarinic receptor suggest a direct dimeric contact mediated by TM5 of each monomer[28]. The potential involvement of the alternative TM1–TM2–H8 (where H8 is helix 8) interface in GPCR oligomerization has previously been indicated by several different biochemical studies[25] and, more recently, by the structure of opsin (Protein Data Bank (PDB) accession 3CAP)[29]. In the case of opioid receptors, it has been shown that a μ-OR TM1 domain fused to a polybasic TAT sequence could disrupt the μ-OR–δ-OR interaction in the mouse spinal cord, resulting in an enhancement of morphine analgesia and a reduction in morphine tolerance[30].

The more prominent interface observed in the μ-OR crystal structure is comprised of TM5 and TM6 of each protomer arranged in a four-helix bundle motif (Fig. 5a). This interface is formed by an extensive network of interactions involving 28 residues in TM5 and TM6 (Fig. 5c and Supplementary Fig. 8). These surface packing interactions are highly complementary and are maintained all along the receptor membrane plane from the extracellular to the intracellular side of the μ-OR (Fig. 5c, d). The $T279^{6.34}$ residue described earlier as having a role in maintaining the receptor in an inactive state is also part of the dimer interface, with the methyl of the threonine contacting $I256^{5.62}$ of the adjacent protomer. It is thus tempting to speculate that dimerization of the μ-OR could have a role in regulating receptor signalling.

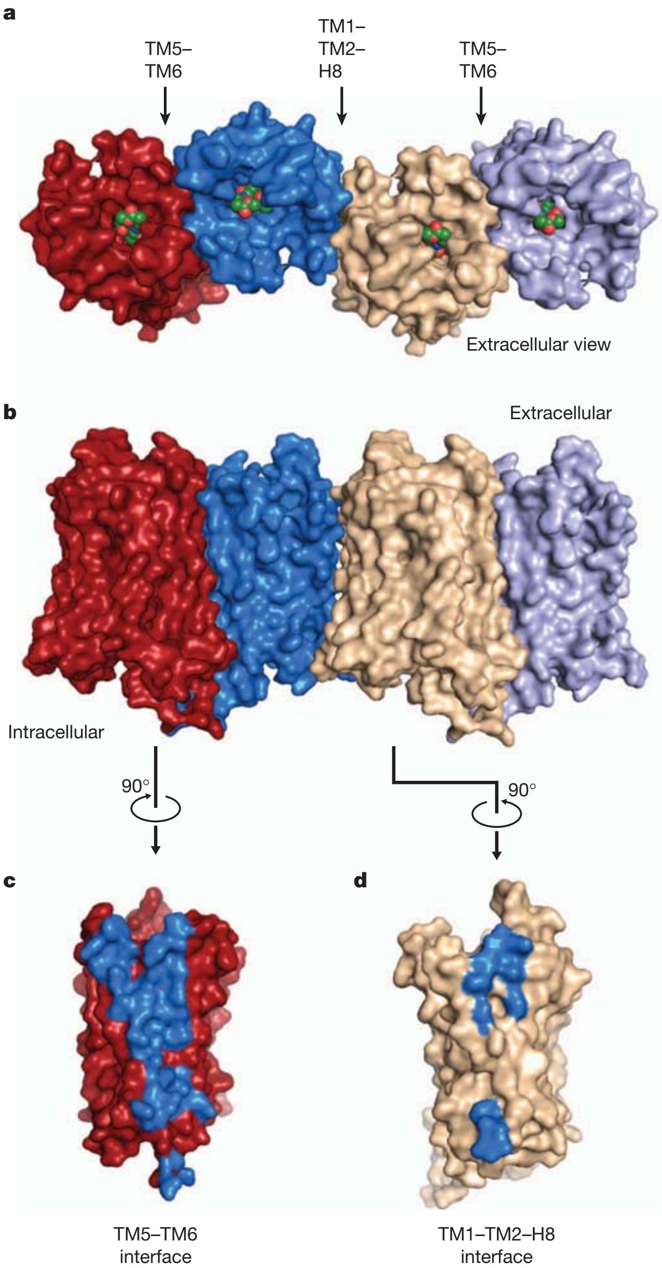


Figure 4 | μ-OR oligomeric arrangement. **a**, **b**, μ-OR crystallized as intimately associated pairs, with two different interfaces as defined in the text. **c**, **d**, The interface defined by TM5 and TM6 (**c**) is much more extensive than for the one defined by TM1–TM2–H8 (**d**).

The observed dimer is of interest because of existing evidence for both homo- and heterodimers (or oligomers) involving the μ-OR[31]. It has been suggested that opioid agonists such as DAMGO and methadone reduce tolerance to morphine *in vivo* by facilitating morphine-induced endocytosis through μ-OR oligomerization[32,33]. These studies implicate allosteric interactions between a protomer bound to DAMGO or methadone and an adjacent protomer bound to morphine. Co-expressing μ-OR and δ-OR in cells results in pharmacological profiles distinct from either receptor expressed alone[34]. Of interest, morphine is more efficacious in cells expressing both μ-OR and δ-OR in the presence of a δ-OR-selective antagonist, suggesting an allosteric interaction between μ-OR and δ-OR protomers[35]. Hetero-oligomerization between μ-OR and non-opioid receptors has

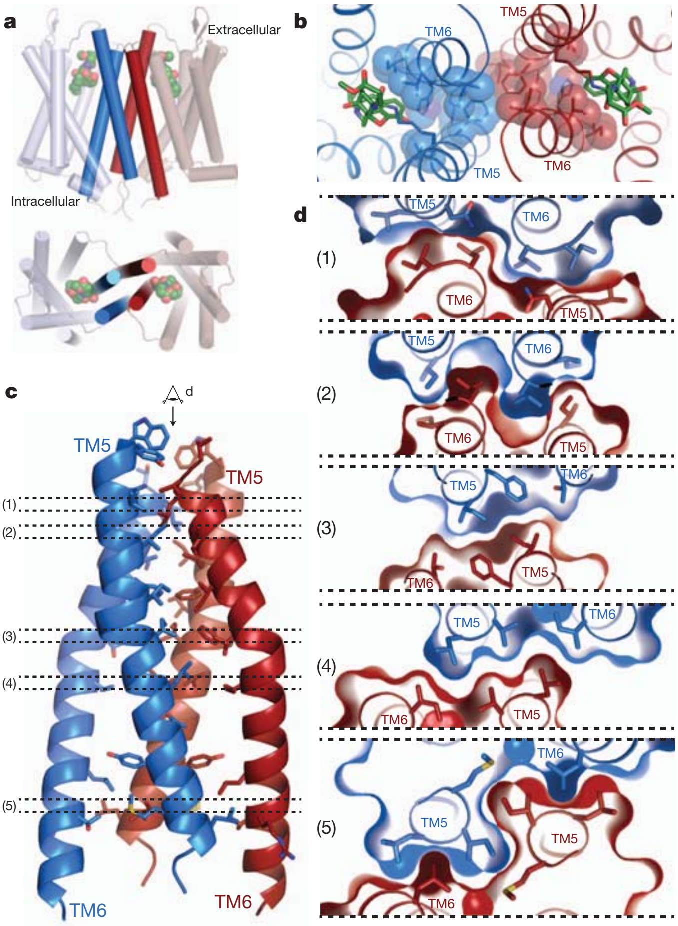


Figure 5 | The four-helix bundle interface. a, Schematic showing the four-helix bundle architecture of the TM5–TM6 interface. **b**, Viewed from the extracellular surface, the binding pocket shows tight association between the ligand (green sticks) and residues that are involved directly or indirectly in forming the dimeric interface (blue spheres). **c**, The four-helix bundle is expanded and shown in detail with interacting residues within 4.2 Å shown as sticks. **d**, Tomographic representation along the dimer interface viewed from the extracellular side (as indicated in panel **c**) showing the high surface complementarity within the four-helix bundle interface.

also been reported[23]. For example, the α_{2a} adrenergic receptor was shown to modulate receptor μ-OR structure and signalling[36].

Consistent with a role for oligomerization in μ-OR function, we observed that the amino acids involved in the dimer interface display a high degree of homology with the δ-OR (Supplementary Figs 9 and 10). Replacing the residues of μ-OR with the corresponding residues from δ-OR would not be predicted to interfere with dimer formation (Supplementary Figs 9 and 10). This analysis also suggests that a μ-OR–δ-OR dimer could share the same interface. Interestingly, in the μ-OR TM5–TM6 dimer, the two binding sites are coupled through a network of packing interactions at the dimeric interface (Fig. 5b). This network could provide a structural explanation for the distinct pharmacological profiles obtained for μ-OR heterodimers and for the allosteric effects of one protomer on the pharmacological properties of the other. This dimeric interface thus provides potential insights into the mechanism of allosteric regulation of one GPCR protomer by the other.

Parallel dimers have also been observed in other GPCR crystal structures, most notably in CXCR4–T4L[18]. Interestingly, the CXCR4 dimer is also related by a two-fold rotational symmetry axis with a receptor arrangement similar but not identical to that seen in μ-OR (Supplementary Fig. 8). However, for the five different CXCR4–T4L crystal structures, the largest calculated contact area between the two CXCR4 protomers is smaller (1,077 Å^2 for PDB accession 3OE0) than in the μ-OR structure (Supplementary Fig. 7), and it presents a comparatively less extensive network of interactions (Supplementary Fig. 8).

The dimeric arrangement of μ-OR across the TM5–TM6 interface observed in the crystal structure would probably preclude either protomer from coupling to G proteins. This is based on structural changes in TM5 and TM6 observed in the recent crystal structure of the β_2-AR–G_s complex[37]. This is also consistent with the observation that inverse agonists stabilize β_2-AR oligomers, while the G protein G_s reduced the extent of oligomerization[38]. However, we were able to model an active structure of μ-OR in complex with G protein based on the crystal structure of the β_2-AR–G_s complex. Here, we observed that a tetramer formed by the association of two dimers through a TM5–TM6 interface would accommodate two G proteins in interaction with the two distal protomers (Supplementary Fig. 11). This model of an activated μ-OR–G-protein oligomeric complex is highly speculative but is compatible with results from a recent biophysical study suggesting that the G-protein G_i remains associated with a μ-OR tetramer stabilized by the agonist morphine[39].

The μ-OR is perhaps the most economically important GPCR in terms of the combined legal and illicit drug market. Although there are a number of effective drugs targeting the μ-OR on the market, the ideal agonist has yet to be developed. The structure of the μ-OR presented here provides the first high-resolution insight, to our knowledge, into a peptide receptor that can also be activated by small-molecule agonist ligands, some of which are the oldest used drugs in human history. This structure will enable the application of structure-based approaches to complement more conventional drug discovery programs. In addition, it may provide novel insights into the role of oligomerization in GPCR function.

METHODS SUMMARY

The μ-OR–T4L fusion protein was expressed in Sf9 insect cells and purified by nickel affinity chromatography followed by Flag antibody affinity chromatography and size-exclusion chromatography. It was crystallized using the lipidic cubic phase technique, and diffraction data were collected at GM/CA-CAT beamline 23ID-D at the Advanced Photon Source at Argonne National Laboratory. The structure was solved by molecular replacement using merged data from 25 crystals.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 December 2011; accepted 9 February 2012.
Published online 21 March 2012.

1. Katzung, B. G. *Basic and Clinical Pharmacology* 10th edn (LANGE McGraw Hill Medical, 2007).
2. Matthes, H. W. *et al.* Loss of morphine-induced analgesia, reward effect and withdrawal symptoms in mice lacking the μ-opioid-receptor gene. *Nature* **383,** 819–823 (1996).
3. Lord, J. A., Waterfield, A. A., Hughes, J. & Kosterlitz, H. W. Endogenous opioid peptides: multiple agonists and receptors. *Nature* **267,** 495–499 (1977).
4. Raffa, R. B., Martinez, R. P. & Connelly, C. D. G-protein antisense oligodeoxyribonucleotides and μ-opioid supraspinal antinociception. *Eur. J. Pharmacol.* **258,** R5–R7 (1994).
5. Shukla, A. K., Xiao, K. & Lefkowitz, R. J. Emerging paradigms of β-arrestin-dependent seven transmembrane receptor signaling. *Trends Biochem. Sci.* **36,** 457–469 (2011).
6. Molinari, P. *et al.* Morphine-like opiates selectively antagonize receptor-arrestin interactions. *J. Biol. Chem.* **285,** 12522–12535 (2010).
7. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into β_2-adrenergic receptor function. *Science* **318,** 1266–1273 (2007).
8. Ballesteros, J. A. & Weinstein, H. *Integrated Methods for the Construction of Three Dimensional Models and Computational Probing of Structure Function Relations in G Protein-Coupled Receptors* Vol. 25 366–428 (Academic, 1995).
9. Chen, C. *et al.* Determination of the amino acid residue involved in [^{3}H]β-funaltrexamine covalent binding in the cloned rat μ-opioid receptor. *J. Biol. Chem.* **271,** 21422–21429 (1996).
10. Huang, P. *et al.* Functional role of a conserved motif in TM6 of the rat μ opioid receptor: constitutively active and inactive receptors result from substitutions of Thr6.34(279) with Lys and Asp. *Biochemistry* **40,** 13501–13509 (2001).

11. Haga, K. *et al.* Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature* **482,** 547–551 (2012).
12. Kruse, A. C. *et al.* Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **482,** 552–556 (2012).
13. Disse, B. *et al.* Ba 679 BR, a novel long-acting anticholinergic bronchodilator. *Life Sci.* **52,** 537–544 (1993).
14. Cassel, J. A., Daubert, J. D. & DeHaven, R. N. [^{3}H]Alvimopan binding to the μ opioid receptor: comparative binding kinetics of opioid antagonists. *Eur. J. Pharmacol.* **520,** 29–36 (2005).
15. Kurowski, M., Rosenbaum, J. S., Perry, D. C. & Sadee, W. [^{3}H]-etorphine and [^{3}H]-diprenorphine receptor binding *in vitro* and *in vivo*: differential effect of Na^+ and guanylyl imidodiphosphate. *Brain Res.* **249,** 345–352 (1982).
16. Sporer, K. A. Acute heroin overdose. *Ann. Intern. Med.* **130,** 584–590 (1999).
17. Alford, B. T., Burkhart, R. L. & Johnson, W. P. Etorphine and diprenorphine as immobilizing and reversing agents in captive and free-ranging mammals. *J. Am. Vet. Med. Assoc.* **164,** 702–705 (1974).
18. Wu, B. *et al.* Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* **330,** 1066–1071 (2010).
19. Mansour, A. *et al.* Key residues defining the μ-opioid receptor binding pocket: a site-directed mutagenesis study. *J. Neurochem.* **68,** 344–353 (1997).
20. Bonner, G., Meng, F. & Akil, H. Selectivity of μ-opioid receptor determined by interfacial residues near third extracellular loop. *Eur. J. Pharmacol.* **403,** 37–44 (2000).
21. Zadina, J. E., Hackler, L., Ge, L. J. & Kastin, A. J. A potent and selective endogenous agonist for the μ-opiate receptor. *Nature* **386,** 499–502 (1997).
22. Seki, T. *et al.* DAMGO recognizes four residues in the third extracellular loop to discriminate between μ- and κ-opioid receptors. *Eur. J. Pharmacol.* **350,** 301–310 (1998).
23. Rozenfeld, R., Gomes, I. & Devi, L. in *The Opiate Receptors* Vol. 23 (ed. Pasternak, G. W.) Ch. 15 407–437 (Humana, 2011).
24. Johnston, J. M. *et al.* Making structural sense of dimerization interfaces of δ opioid receptor homodimers. *Biochemistry* **50,** 1682–1690 (2011).
25. Fanelli, F. & De Benedetti, P. G. Update 1 of: computational modeling approaches to structure-function analysis of G protein-coupled receptors. *Chem. Rev.* **111,** PR438–PR535 (2011).
26. Hebert, T. E. *et al.* A peptide derived from a β_2-adrenergic receptor transmembrane domain inhibits both receptor dimerization and activation. *J. Biol. Chem.* **271,** 16384–16392 (1996).
27. Granier, S. *et al.* A cyclic peptide mimicking the third intracellular loop of the V2 vasopressin receptor inhibits signaling through its interaction with receptor dimer and G protein. *J. Biol. Chem.* **279,** 50904–50914 (2004).
28. Hu, J. *et al.* Structural aspects of M3 muscarinic acetylcholine receptor dimer formation and activation. *FASEB J.* **26,** 604–616 (2011).
29. Park, J. H., Scheerer, P., Hofmann, K. P., Choe, H. W. & Ernst, O. P. Crystal structure of the ligand-free G-protein-coupled receptor opsin. *Nature* **454,** 183–187 (2008).
30. He, S. Q. *et al.* Facilitation of μ-opioid receptor activity by preventing δ-opioid receptor-mediated codegradation. *Neuron* **69,** 120–131 (2011).
31. Jordan, B. A. & Devi, L. A. G-protein-coupled receptor heterodimerization modulates receptor function. *Nature* **399,** 697–700 (1999).
32. He, L., Fong, J., von Zastrow, M. & Whistler, J. L. Regulation of opioid receptor trafficking and morphine tolerance by receptor oligomerization. *Cell* **108,** 271–282 (2002).
33. He, L. & Whistler, J. L. An opiate cocktail that reduces morphine tolerance and dependence. *Curr. Biol.* **15,** 1028–1033 (2005).
34. George, S. R. *et al.* Oligomerization of μ- and δ-opioid receptors. Generation of novel functional properties. *J. Biol. Chem.* **275,** 26128–26135 (2000).
35. Gomes, I., Ijzerman, A. P., Ye, K., Maillet, E. L. & Devi, L. A. G protein-coupled receptor heteromerization: a role in allosteric modulation of ligand binding. *Mol. Pharmacol.* **79,** 1044–1052 (2011).
36. Vilardaga, J. P. *et al.* Conformational cross-talk between α_{2A}-adrenergic and μ-opioid receptors controls cell signaling. *Nature Chem. Biol.* **4,** 126–131 (2008).
37. Rasmussen, S. G. *et al.* Crystal structure of the β_2 adrenergic receptor–G_s protein complex. *Nature* **477,** 549–555 (2011).
38. Fung, J. J. *et al.* Ligand-regulated oligomerization of β_2-adrenoceptors in a model lipid bilayer. *EMBO J.* **28,** 3315–3328 (2009).
39. Golebiewska, U., Johnston, J. M., Devi, L., Filizola, M. & Scarlata, S. Differential response to morphine of the oligomeric state of μ-opioid in the presence of δ-opioid receptors. *Biochemistry* **50,** 2829–2837 (2011).
40. Portoghese, P. S., Sultana, M. & Takemori, A. E. Design of peptidomimetic δ opioid receptor antagonists using the message-address concept. *J. Med. Chem.* **33,** 1714–1720 (1990).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge support from INSERM (S.G.), the Stanford Medical Scientist Training Program (A.M.), the National Science Foundation (A.C.K.), the Lundbeck Foundation (J.M.M.), the National Institutes of Health Grants NS028471 (B.K.K.) and DA031418 (B.K.K. and R.K.S.), and the Mathers Foundation (B.K.K. and W.I.W.).

Author Contributions A.M., A.C.K. and S.G. designed experiments, performed research and analysed data. T.S.K. and F.S.T. expressed and purified receptor. J.M.M. performed preliminary biochemical experiments with wild-type μ-OR. R.K.S. contributed to the effort of μ-OR crystallization and writing of the manuscript. W.I.W. supervised diffraction data analysis and model refinement. L.P. built the tetramer model and helped with the analysis of the dimer interfaces. A.M., A.C.K., S.G. and B.K.K. prepared the manuscript. S.G. and B.K.K. supervised the research.

Author Information Coordinates and structure factors for μ-OR–T4L are deposited in the Protein Data Bank under accession code 4DKL. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.G. (granier@stanford.edu) or B.K.K. (kobilka@stanford.edu).

METHODS

Expression and purification. Previously crystallized GPCRs show little density for the poorly ordered amino- and carboxy-terminal domains. Although these domains are not critical for maintaining high ligand affinity, these flexible regions may inhibit crystallogenesis[7]. We therefore removed these regions in the receptor construct used for crystallography. Specifically, a TEV protease recognition site was introduced after reside G51 in the amino terminus and the C terminus was truncated after Q360. The short third intracellular loop of μ-OR, consisting of residues 264–269, was replaced with T4L residues 2–161 in a manner described previously[7]. To facilitate receptor purification, a Flag M1 tag was added to the N terminus and an octa-histidine tag was appended to the C terminus. Finally, a proline residue was introduced N-terminal to the octahistidine tag to allow efficient removal of C-terminal histidines by carboxypeptidase A. For these studies, we used the *M. musculus* μ-OR sequence because it is expressed at higher levels. The mouse and human μ-OR share 94% sequence identity and there are only four residues in the resolved part of the structure that differ between the mouse and human μ-OR. These include residues 66, 137, 187 and 306, which are all in the extracellular or intracellular loops of μ-OR and do not make contacts in the ligand-binding pocket. The final crystallization construct (μ-OR–T4L) is shown in a representative snake diagram in Supplementary Fig. 1a.

We compared the pharmacological properties of μ-OR–T4L to those of the wild-type receptor (Supplementary Fig. 1b). Both constructs showed identical affinity for the radiolabelled antagonist [^{3}H]-diprenorphine ([^{3}H]DPN).

The μ-OR–T4L construct was expressed in Sf9 cells using the baculovirus system. Culture media was supplemented with 10 μM naloxone to stabilize the receptor during expression. Cells were infected at a density of 4×10^6 cells per ml and culture flasks were shaken at 27 °C for 48 h. After harvesting, cells were lysed by osmotic shock in a buffer comprised of 10 mM Tris-HCl pH 7.5, 1 mM EDTA, 100 μM TCEP, 1 μM naloxone and 2 mg ml^{-1} iodoacetamide to block reactive cysteines. Extraction of μ-OR–T4L from Sf9 membranes was done with a Dounce homogenizer in a solubilization buffer comprised of 0.5% dodecyl maltoside (DDM), 0.3% 3-[(3-Cholamidopropyl) dimethylammonio]-1-propanesulphonate (CHAPS), 0.03% cholesterol hemisuccinate (CHS), 20 mM HEPES pH 7.5, 0.5 M NaCl, 30% v/v glycerol, 2 mg ml^{-1} iodoacetamide, 100 μM TCEP and 1 μM naloxone. After centrifugation, nickel-NTA agarose was added to the supernatant, stirred for 2 h, and then washed in batch with 100*g* spins for 5 min each with a washing buffer of 0.1% DDM, 0.03% CHAPS, 0.01% CHS, 20 mM HEPES pH 7.5 and 0.5 M NaCl. The resin was poured into a glass column and bound receptor was eluted in washing buffer supplemented with 300 mM imidazole.

We used anti-Flag M1 affinity resin to purify μ-OR–T4L further and to exchange the ligand with the covalent antagonist β-FNA. Nickel-resin eluate was loaded onto anti-Flag M1 resin and washed extensively in the presence of 10 μM β-FNA. The detergent DDM was then gradually exchanged over 1 h into a buffer with 0.01% lauryl maltose neopentyl glycol (MNG) and the NaCl concentration was lowered to 100 mM. Receptor was eluted from the anti-Flag M1 affinity resin with 0.2 mg ml^{-1} Flag peptide and 5 mM EDTA in the presence of 1 μM β-FNA. To remove the N terminus of μ-OR–T4L, TEV protease was added at 1:3 w/w (TEV:μ-OR–T4L) and incubated at room temperature (23 °C) for 1 h. Receptor was then treated with carboxypeptidase A (1:100 w/w) and incubated overnight at 4 °C to remove the octa-histidine tag. The final purification step separated TEV and carboxypeptidase A from receptor by size exclusion chromatography on a Sephadex S200 column (GE Healthcare) in a buffer of 0.01% MNG, 0.001% CHS, 100 mM NaCl, 20 mM HEPES pH 7.5 and 1 μM β-FNA. After size exclusion, β-FNA was added to a final concentration of 10 μM. The resulting receptor preparation was pure and monodisperse (Supplementary Fig. 12).

Crystallization and data collection. Purified μ-OR–T4L receptor was concentrated to 30 mg ml^{-1} using a Vivaspin sample concentrator with a 50 kDa molecular weight cut-off (GE Healthcare) and crystallization was performed using the *in meso* method[41]. Concentrated μ-OR–T4L was reconstituted into 10:1 monoolein:cholesterol (Sigma) in a ratio of 1:1.5 parts by weight receptor:lipid mixture. Reconstitution was done by the two-syringe method[41]. The resulting mesophase was dispensed onto glass plates in 80-nl drops and overlaid with 700 nl precipitant solution by a Gryphon LCP robot (Art Robbins Instruments). Crystals grew in precipitant solution consisting of 30–38% PEG 400, 100 mM HEPES pH 7.0, 7.5% DMSO and 300 mM lithium sulphate. Crystals were observed after 24 h and grew to full size after 5 days. Typical crystals before harvesting are shown in Supplementary Fig. 2.

Diffraction data were collected at Advanced Photon Source GM/CA-CAT beamline 23ID-D using a beam size of 10 μm. Owing to radiation damage, the diffraction quality decayed during exposure. Wedges of 10–20 degrees were collected and merged from 25 crystals using HKL2000[42]. Diffraction quality ranged from 2.4–3.5 Å in most cases. The structure of the μ-OR was solved by molecular replacement in Phaser[43] using the CXCR4 receptor as a search model. We improved the initial model by iteratively building regions of the receptor in Coot[44] and refining in Phenix[45]. We used translation libration screw-motion (TLS) refinement with groups generated within Phenix. Electron density suggested the presence of a cholesterol molecule and a monoolein lipid within the lipidic layer. These were subsequently incorporated into the model. To assess the overall quality of the final structure, we used MolProbity[46]. The resulting statistics for data collection and refinement are shown in Supplementary Table 1. Figures were prepared in PyMOL[47].

Saturation binding experiments. Membrane homogenates were prepared from Sf9 cells expressing either wild-type μ-OR or μ-OR–T4L. Membranes containing μ-OR or μ-OR–T4L were incubated with the opioid antagonist, [^{3}H]DPN for 1 h at 22 °C in 0.5 ml of binding buffer containing 75 mM Tris-HCl pH 7.4, 1 mM EDTA, 5 mM $MgCl_2$, 100 mM NaCl. To determine the affinity for diprenorphine, we used [^{3}H]DPN concentrations ranging from 0.1 to 13.5 nM. High concentrations of un-labelled naloxone (1 μM) were used to determine non-specific binding. To separate unbound [^{3}H]-ligand, binding reactions were rapidly filtered over GF/C Brandel filters. The filters were then washed three times with 5 ml ice-cold binding buffer. Radioactivity was assayed by liquid scintillation counting. The resulting data were analysed using Prism 5.0 (GraphPad Software). [^{3}H]DPN (specific activity: 55.0 Ci mmol^{-1}) was obtained from PerkinElmer Life Sciences.

41. Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nature Protocols* **4,** 706–731 (2009).
42. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
43. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Cryst.* **40,** 658–674 (2007).
44. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
45. Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. A robust bulk-solvent correction and anisotropic scaling procedure. *Acta Crystallogr. D* **61,** 850–855 (2005).
46. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
47. Schrodinger, L. The PyMOL Molecular Graphics System v.1.3r1. (2010).

LETTER

doi:10.1038/nature11022

Earth-like sand fluxes on Mars

N. T. Bridges[1], F. Ayoub[2], J-P. Avouac[2], S. Leprince[2], A. Lucas[2] & S. Mattson[3]

Strong and sustained winds on Mars have been considered rare, on the basis of surface meteorology measurements and global circulation models[1,2], raising the question of whether the abundant dunes and evidence for wind erosion seen on the planet are a current process. Recent studies[3–6] showed sand activity, but could not determine whether entire dunes were moving—implying large sand fluxes—or whether more localized and surficial changes had occurred. Here we present measurements of the migration rate of sand ripples and dune lee fronts at the Nili Patera dune field. We show that the dunes are near steady state, with their entire volumes composed of mobile sand. The dunes have unexpectedly high sand fluxes, similar, for example, to those in Victoria Valley, Antarctica, implying that rates of landscape modification on Mars and Earth are similar.

The Martian surface displays abundant bedforms (ripples and dunes)[7], and evidence for wind erosion ranging from centimetre-scale ventifact rock textures[8] to kilometre-scale yardangs and exhumed mantles[7]. But whether these features are actively moving has been an open question, as sand-transporting winds have been considered rare in the low-density atmosphere of Mars[1,2]. Although many bedforms have been interpreted as static, relict features[7], and dune formation times are predicted to be five orders of magnitude slower than on Earth[9], recent high-resolution orbiter[3–6] and rover[10] images have provided evidence of sand movement. Whether these observations document surficial migration, or bedforms in equilibrium (the full volume undergoing movement)— and therefore large sand fluxes capable of actively eroding the surface—could not be determined using the traditional measurement techniques of the earlier studies. Resolution of this problem has implications for understanding past Martian climates, as it has been suggested that significant erosion on Mars may have required a higher-pressure atmosphere in the past[11].

Recent advances in optical image correlation[12] have allowed dune migration rates and associated sand fluxes in terrestrial dune fields to be estimated from satellite data[13]. We took advantage of the High Resolution Imaging Science Experiment (HiRISE) on the Mars Reconnaissance Orbiter, a push-broom imager with pixel sizes of ~25 cm[14], to track the displacement of sand ripples covering the dunes. We implemented the HiRISE geometry in the "Co-registration of Optically Sensed Images and Correlation" (COSI-Corr) tool suite, which provides quantitative surface dynamics measurements by automatic and precise orthorectification, co-registration, and subpixel correlation of images[12]. The resulting data rival those obtained using remotely sensed images of Earth dunes[13].

The high spatial resolution of HiRISE combined with COSI-Corr allows us to quantify dune ripple migration rates and the derivation of sand flux across the entire image, a critical measurement that we perform for the first time on a planetary surface. To undertake this investigation, we chose the Nili Patera dune field, a location containing abundant barchan dunes with morphology typical of dunes elsewhere on Mars and on Earth, and for which localized ripple migration has been identified using visual comparison of images acquired at different times[3,6] (Supplementary Fig. 1). Four HiRISE images centred on the Nili Patera dune field (8.8° N, 67.3° E) were used: two to quantify changes that occurred in the time interval (105 Earth days) between their acquisition (subsequently referred to as images T1 and T2); and another two to construct a stereo-derived Digital Elevation Model, on which the images were orthorectified and co-registered (images S1 and S2; Supplementary Table 1). The correlation of T1 and T2 provides dense measurements of ripple migration (Fig. 1a). We find that ripple migration occurs across the entire dune field, and increases with elevation along the stoss slopes. Measurable ripple motion is up to 4.5 m, confirming high sand activity. As dunes become more sheltered towards the southwest, ripple displacement generally decreases, consistent with active, northeasterly winds. The azimuthal distribution of displacement vectors (Fig. 1a, inset) is consistent with the southwesterly trend inferred from the orientation of the dune slip faces and barchan horns.

Ripple displacement (d_r) increases linearly with elevation on a given dune (h_D) (Fig. 2). The fastest ripples, those with the steepest slopes in Fig. 2, moved so far that the correlation breaks down once the displacement exceeds a distance approximately equal to the ripple wavelength (4.6 ± 0.09 m; see Supplementary Information). This linear relationship is consistent with the behaviour of steady-state migrating dunes in which mass is conserved while shape and volume are maintained[15], and wind shear stress increases with dune elevation[9,16] (see Supplementary Information).

We also measured dune migration from the advancement of lee fronts between T2 and S1 (941 days), taking advantage of the accurate registration of the two images obtained with COSI-Corr. Fronts showed measurable advancement (Supplementary Animation), but were negligible between T1 and T2 (105 d) except for a few isolated cases where avalanches occurred[3,6]. These measurements were performed only where the lee advance was clear, and therefore may not represent the average migration of the whole dune field.

Our measurements of ripples and dune migration are related, as they both reflect sand transport, and can be used to estimate sand flux. Sand transport results from saltation and reptation[17,18]. Saltation is the hopping motion of grains over long trajectories which, when they collide with a sand bed, results in a splash of shorter reptation trajectories. Reptation causes ripple migration, whereas both processes contribute to dune advancement. The reptation sand flux is estimated by multiplying the ripples' migration rate by their average height, h_r, estimated to be 20 ± 6 cm (Supplementary Information). Assuming the reptation sand flux equals that of the whole dune, the dune migration rate is $d_r h_r/(h_D t)$ (Supplementary Information), where t is the time interval (105 d). The histogram of dune migration rates over the whole study area peaks at an average value of ~0.1 m yr^{-1} (Earth year) (Fig. 3). This distribution is consistent with the dune speeds derived from linear fits to the selected ripple profiles of Fig. 2, which range from 0.03 ± 0.01 to 0.27 ± 0.08 m yr^{-1} (Fig. 3). The relative uncertainty in dune migration rate derived from these measurements is estimated to be less than 20% (Fig. 3, inset table); however, this is a minimum estimate because the saltation sand flux contribution is not yet considered. Extrapolating the dune migration rates derived from lee-front advancement to a year shows that the lee-derived rates are approximately five times larger than the ripple-derived rates (Fig. 3).

[1]Space Department, Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland 20723, USA. [2]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA. [3]Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA.

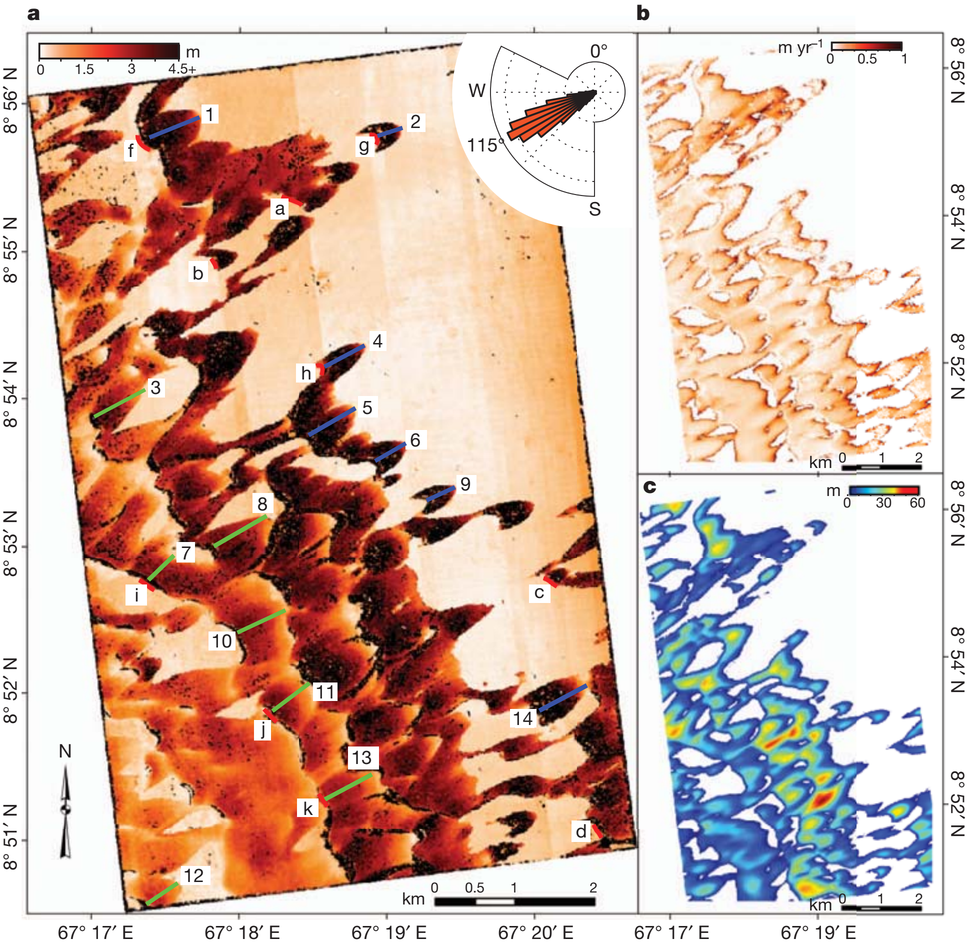


Figure 1 | Ripple migration, dune migration and dune elevation. **a**, Ripple displacements in the Nili Patera dune field derived from correlating HiRISE images PSP_004339_1890 (30 June 2007) and PSP_005684_1890 (13 October 2007). Green and blue numbered lines show where profiles of ripple displacement were retrieved. Red lines with letters show where dune lee-front displacements were measured between images PSP_004339_1890 and ESP_017762_1890 (11 May 2010). Inset rose diagram shows distribution of ripple migration azimuth. **b**, Dune migration rate derived from the ripple migration rates and dune elevation. **c**, Dune elevation relative to bedrock base. Elevation and height maps are based on stereo images ESP_017762_1890 and ESP_018039_1890 (2 June 2010). See Supplementary Information for details of how dune height was estimated.

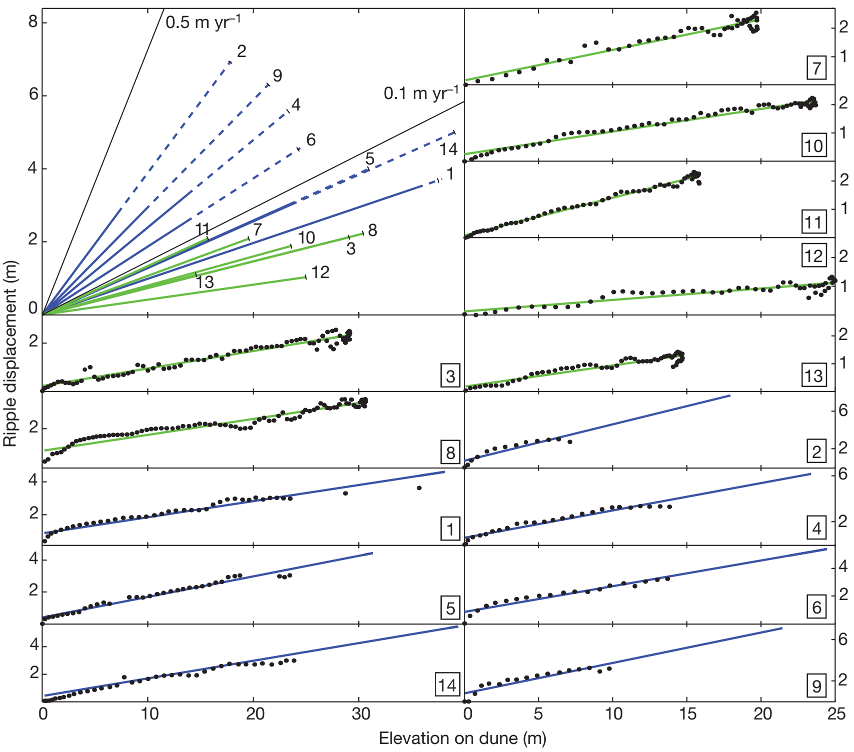


Figure 2 | Linear correlation between ripple migration and dune height. Upper left frame: ripple displacement versus local dune elevation for profiles on upwind (blue) and downwind (green) dunes. Profile locations are shown in Fig. 1a. ('Upwind' dunes are in the northeast part of the field, unsheltered from prevailing northeasterly winds; 'downwind' dunes are in the southwest part of the field, partially sheltered by the dunes to the northeast.) Solid lines are best-fitting linear functions, shifted to go through the origin to facilitate slope comparisons. Dashed lines are extrapolations out to the dune crest for cases where displacements could not be estimated, owing to decorrelation of the ripple patterns. Black lines are isopleths of dune migration rates. The individual profiles and measured ripple displacements are shown in the other frames.

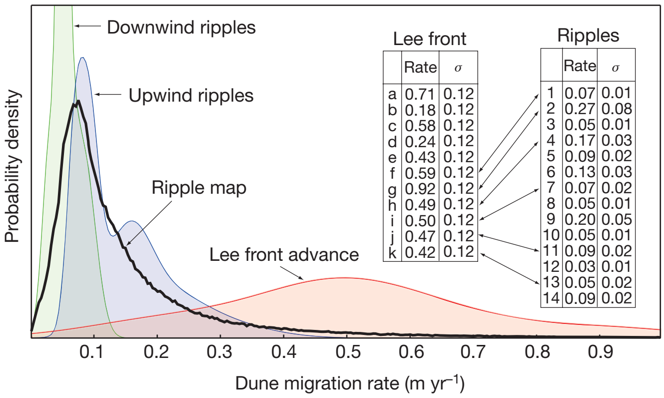


Figure 3 | Dune migration rates. Normalized histogram (black) of dune migration rates over the 105-day T1–T2 time interval, derived from all measurements of Fig. 1b where dune height (elevation above the bedrock base) exceeds 0.5 m. (All time intervals and rates in this paper refer to Earth days and years.) The right-side tail is mostly attributable to rates measured on the dunes' upwind and lateral edges, where the measured rates are probably erroneously large because the ripples there are not fully developed to the equilibrium height that is assumed in the calculations. (See Supplementary Information.) These large values appear as rims around the dunes in Fig. 1b. The blue and green probability density functions show, respectively, the migration rates inferred for individual upwind and downwind dunes (Fig. 2). (See inset tables and Supplementary Table 2 for all migration rates and uncertainties.) The mean heights of the upwind and downwind dunes are 27 and 22 m, respectively. The red probability density function shows migration rates derived from lee-front advance over the 941-day T2–S1 interval. The location of these lee fronts is shown in Fig. 1a (except for measurement e, which is located outside the area shown in the figure). Black arrows between the inset tables link dune migration rates obtained on the same dune from lee-front tracking (thus resulting from reptation and saltation) and from ripple migration (resulting from reptation only). Assumptions about ripple geometry result in uncertainty in the mean height of the ripples (see Supplementary Information), which propagates to a 20% uncertainty in dune migration rate.

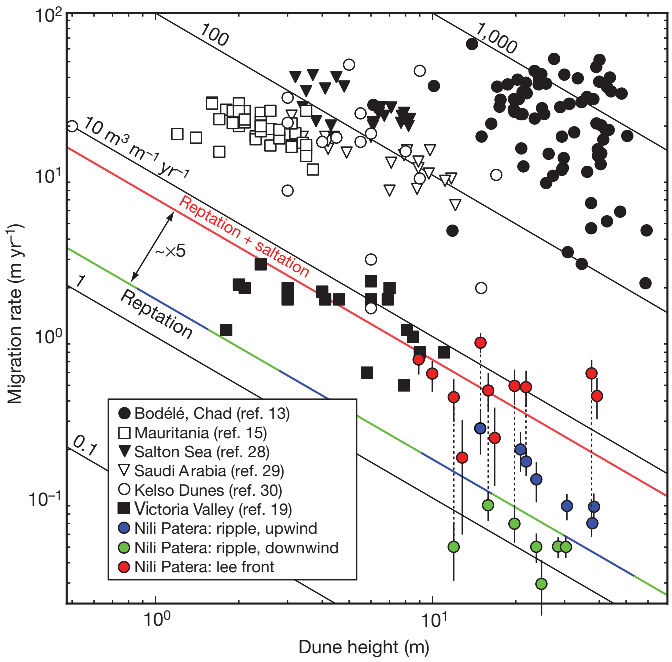


Figure 4 | Comparison of dune migration rates and sand flux on Mars and Earth. Dune migration rates versus dune height for a number of sites on Earth (with reference numbers in parentheses); the 14 dunes selected in Fig. 2; and dunes for which the lee-front advance was measured (locations in Fig. 1a). Black diagonal lines are isopleths of sand flux. Red and blue/green diagonal lines are mean sand fluxes derived from the lee-front advance and ripple migration measurements, respectively. Vertical error bars show 1σ (1 standard deviation) confidence intervals for the dune migration rates. The mean sand flux derived from the lee-front advance is 6.9 ± 0.52 (1σ) $m^3 m^{-1} yr^{-1}$, and the mean sand flux derived from the ripple migration measurements is 1.4 ± 0.08 $m^3 m^{-1} yr^{-1}$. (See Supplementary Table 2 for individual measurements.) The factor of five between the two fluxes suggests that the saltation flux is about four times the reptation flux.

The higher values reflect the contribution of the saltation sand flux, which is not considered in the ripple-derived rates.

Comparison with terrestrial dunes (Fig. 4), shows that the Nili migration rates are about 1–2 orders of magnitude slower than for dunes of comparable height on Earth. Multiplying the dune migration rates by their maximum heights (h_{Dmax}) gives sand fluxes at the dune crests: $Q_0 = d_D h_{Dmax}/t$, where d_D is dune displacement[13]. Mean fluxes for reptation and reptation plus saltation are 1.4 and 6.9 $m^3 m^{-1} yr^{-1}$, respectively (Fig. 4). This is comparable to sand fluxes for dunes in Victoria Valley, Antarctica[19]. Terrestrial studies show that bulk and interdune sand fluxes are about one-third of the crest flux[15], so that typical fluxes in Nili should be ~2.3 $m^3 m^{-1} yr^{-1}$. The Nili dunes have ~1,000 times the volume of those in Victoria Valley, yet similar sand fluxes, indicating that the characteristic timescales of formation are ~1,000 times longer, showing that dunes on Mars evolve much more slowly than their counterparts on Earth. The timescale associated with the formation and evolution of the Nili Patera dune field, estimated by dividing the dunes' volume by the average sand flux times the length scale of the dune field, is ~9,800 yr. Similarly, turnover times needed for dunes to migrate over a distance equal to their length are very short, ranging from ~170 yr for the fastest dune (Fig. 1a, dune c) to a few thousand years for the slower ones.

The observed correlation between ripple and dune displacement implies that the entire volume of the dunes is composed of mobile sand. The alternative, a mobile rippled skin over an indurated sand core, cannot be maintained for long time periods and is therefore improbable. The measured ripple reptation flux implies a rapid erosion rate of ~0.01 m yr^{-1} (mean dune speed of 0.1 m yr^{-1}, multiplied by the tangent of the average stoss slope (6°)). In the absence of a net influx from saltation to compensate for this erosion, the dunes would erode very rapidly. Because saltation drives reptation, the two processes scale[17,18], so that saltation flux should increase up the stoss slope in proportion to the reptation flux. This implies dune erosion as great as 0.05 m yr^{-1}. Such a rapid rate would quickly erode any mobile sand layer and expose the indurated core, effectively shutting off subsequent ripple migration. In this picture, the Nili ripple migration rates and patterns would represent a very short time window (maximum duration of the order of the turnover time—a few hundred to a few thousand years) following the formation of an immobile dune core. Finding the dunes in this rare state seems very improbable. More likely is that erosion of the stoss slope is compensated by deposition on the lee front, resulting in whole dune migration and complete recycling of the entire dune volume.

These results demonstrate that conditions in Nili Patera, and probably over much of Mars, are sufficient to move large dunes and transport fluxes of sand equivalent to those on Earth. This is in contrast to predictions from the Ames General Circulation Model (GCM)[2] that threshold wind speeds sufficient to move sand at Nili Patera should not occur. The spatial resolution of GCMs is insufficient to resolve boundary-layer turbulence that may cause gusts above threshold[20]. Even mesoscale simulations need a spatial resolution of kilometres to a few hundred metres to properly model the atmospheric turbulence that accounts for topography and thermal contrasts at the scales of individual dunes[21,22]. The work exemplified in this study can be applied to other regions of Mars, thereby providing ground calibrations for GCMs and mesoscale models, and descriptions of small-scale atmospheric turbulence.

The occurrence of such large sand fluxes, despite the limited winds in Mars' low-density atmosphere[1,2], is probably related to fundamental differences in how sand is mobilized by the wind on Mars compared to

Earth. Although saltation due to aerodynamic shear at fluid threshold is required to initiate grain motion, once started, the sand ejection resulting from grain impact is the major contributor to the particle flux. On Earth, this results in the wind speed required to maintain saltation being about 80% that required for initiation[23]. But because of the higher and longer saltation trajectories on Mars, grains are accelerated to a greater fraction of the wind speed than on Earth, resulting in impact threshold speeds that are only about 10% of the fluid threshold, equivalent in magnitude to that on Earth[24]. Thus, once saltation is initiated by low-frequency gusts, moderate wind speeds can maintain significant fluxes of sand.

To assess the derived sand fluxes in regard to landscape modification, we consider the abrasion susceptibility, S_a, defined as the mass of sand required to erode a unit mass of rock. For basalt grains striking basaltic rocks at the impact threshold for Mars, $S_a = 2 \times 10^{-6}$ (ref. 25). The average flux of $2.3\,m^3\,m^{-1}\,yr^{-1}$ implies that for the saltation trajectories of 0.1–0.5 m that are likely on Mars[24], the abrasion rate would be $\sim$1–10 μm yr^{-1} on flat ground, and $\sim$10–50 μm yr^{-1} for a vertical rock face (see Supplementary Information), spanning field measurements of basalt abrasion rates in Victoria Valley of $\sim$30–50 μm yr^{-1} (ref. 26).

One view of Mars has been that conditions since the end of the Hesperian period (1.8–3.5 Gyr ago) have been fairly static, with very low erosion rates[27]. This study shows that this is not the case at Nili Patera, and probably not at other areas of Mars where there are significant gusts of sand and wind. This may explain why vast areas of the Martian surface show evidence of erosion and removal, including of mantle materials for which the processes and agents of exhumation have been a mystery[7], yet also contain fields of large dunes that migrate at relatively slow rates. Over long time periods, it may be that much or all of Mars has been subjected to large sand fluxes, with associated erosional modification of the landscape. The techniques reported here can be applied to many dunes and other slowly changing features on Mars and Earth, allowing sand flux and landscape modification to be assessed in a variety of terrains, latitudes, seasons and climates.

Received 6 September 2011; accepted 6 March 2012.
Published online 9 May 2012.

1. Arvidson, R. E., Guinness, E. A., Moore, H. J., Tillman, J. & Wall, S. D. Three Mars years: Viking Lander 1 imaging observations. *Science* **222,** 463–468 (1983).
2. Haberle, R. M., Murphy, J. R. & Schaeffer, J. Orbital change experiments with a Mars General Circulation Model. *Icarus* **161,** 66–89 (2003).
3. Silvestro, S., Fenton, L. K., Vaz, D. A., Bridges, N. T. & Ori, G. G. Ripple migration and dune activity on Mars: Evidence for dynamic processes. *Geophys. Res. Lett.* **37,** L20203 (2010).
4. Chojnacki, M., Burr, D. M., Moersch, J. E. & Michaels, T. I. Orbital observations of contemporary dune activity in Endeavour Crater, Meridiani Planum, Mars. *J. Geophys. Res.* **116,** E00F19 (2011).
5. Hansen, C. J. *et al.* Seasonal erosion and restoration of Mars' northern polar dunes. *Science* **331,** 575–578 (2011).
6. Bridges, N. T. *et al.* Planet-wide sand motion on Mars. *Geology* **40,** 31–34 (2012).
7. Malin, M. C. & Edgett, K. S. Mars Global Surveyor Mars Orbiter Camera: Interplanetary cruise through primary mission. *J. Geophys. Res.* **106,** 23,429–23,570 (2001).
8. Laity, J. E. & Bridges, N. T. Ventifacts on Earth and Mars: Analytical, field, and laboratory studies supporting sand abrasion and windward feature development. *Geomorphology* **105,** 202–217 (2009).
9. Claudin, P. & Andreotti, B. A scaling law for aeolian dunes on Mars, Venus, Earth, and for subaqueous ripples. *Earth Planet. Sci. Lett.* **252,** 30–44 (2006).
10. Sullivan, R. *et al.* Wind-driven particle mobility on Mars: Insights from Mars Exploration Rover observations at "El Dorado" and surroundings at Gusev Crater. *J. Geophys. Res.* **113,** E06S07 (2008).
11. Armstrong, J. C. & Leovy, C. B. Long-term wind erosion on Mars. *Icarus* **176,** 57–74 (2005).
12. Leprince, S., Barbot, S., Ayoub, F. & Avouac, J. P. Automatic and precise orthorectification, coregistration, and subpixel correlation of satellite images, application to ground deformation measurements. *IEEE Trans. Geosci. Rem. Sens.* **45,** 1529–1558 (2007).
13. Vermeesch, P. & Drake, N. Remotely sensed dune celerity and sand flux measurements of the world's fastest barchans (Bodélé, Chad). *Geophys. Res. Lett.* **35,** L24404 (2008).
14. McEwen, A. S. *et al.* The High Resolution Imaging Science Experiment (HiRISE) during MRO's Primary Science Phase (PSP). *Icarus* **205,** 2–37 (2010).
15. Ould Ahmedou, D. *et al.* Barchan dune mobility in Mauritania related to dune and interdune sand fluxes. *J. Geophys. Res.* **112,** F02016 (2007).
16. Andreotti, B., Claudin, P. & Pouliquen, O. Aeolian sand ripples: experimental evidence of fully developed states. *Phys. Rev. Lett.* **96,** 028001 (2006).
17. Anderson, R. S. A theoretical model for aeolian impact ripples. *Sedimentology* **34,** 943–956 (1987).
18. Andreotti, B. A two-species model of aeolian sand transport. *J. Fluid Mech.* **510,** 47–70 (2004).
19. Bourke, M. C., Ewing, R. C., Finnegan, D. & McGowan, H. A. Sand dune movement in Victoria Valley, Antarctica. *Geomorphology* **109,** 148–160 (2009).
20. Fenton, L. K. &. M. i. c. h. a. e. l. s. T. J. Characterizing the sensitivity of daytime turbulent activity on Mars with the MRAMS LES: early results. *Mars* **5,** 159–171 (2010).
21. Fenton, L. J., Toigo, A. & Richardson, M. I. Aeolian processes in Proctor Crater on Mars: mesoscale modeling of dune-forming winds. *J. Geophys. Res.* **110,** E06005 (2005).
22. Spiga, A. & Forget, F. A new model to simulate the Martian mesoscale and microscale atmospheric circulation: validation and first results. *J. Geophys. Res.* **114,** E02009 (2009).
23. Bagnold, R. A. *The Physics of Blown Sand and Desert Dunes* (Dover Publications, 1954).
24. Kok, J. F. Difference in the wind speeds required for initiation versus continuation of sand transport on Mars: implications for dunes and dust storms. *Phys. Rev. Lett.* **104,** 074502 (2010).
25. Greeley, R. *et al.* Rate of wind abrasion on Mars. *J. Geophys. Res.* **87,** 10,009–10,024 (1982).
26. Malin, M. C. Rates of geomorphic modification in ice-free areas, southern Victoria Land, Antarctica. *Antarct. J. US* **20,** 18–21 (1986).
27. Golombek, M. P. *et al.* Erosion rates at the Mars Exploration Rover landing sites and long-term climate change on Mars. *J. Geophys. Res.* **111,** E12S10 (2006).
28. Long, J. T. & Sharp, R. S. Barchan-dune movement in Imperial Valley, California. *Geol. Soc. Am. Bull.* **75,** 149–156 (1964).
29. Fryberger, S. *et al.* Wind sedimentation in the Jafurah sand sea, Saudi Arabia. *Sedimentology* **31,** 413–431 (1984).
30. Lancaster, N. Controls on aeolian activity: some new perspectives from the Kelso Dunes, Mojave Desert, California. *J. Arid Environ.* **27,** 113–125 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Discussions with R. Ewing, C. Narteau, and S. Silvestro on bedforms and R. Kirk on stereo data significantly improved this research. This research was supported by grants from NASA's Mars Data Analysis Program, the Keck Institute for Space Studies, and seed funding from the Jet Propulsion Laboratory's Director's Research and Development Fund.

Author Contributions N.T.B. is the principal investigator of the NASA MDAP grant that partially funded this work, chose the study area, defined major science questions, participated in data analysis, compared results to other Mars and terrestrial data, and led the writing of the paper. F.A. processed all the data with COSI-Corr, produced the figures and Supplementary Information ancillary materials, and played a major role in quantitative analysis and text writing. J-P.A. initiated the project, contributed fundamental ideas on sand flux and dune movement, supervised the data analysis and contributed to text writing. S.L. evaluated all COSI-Corr and quantitative results. A.L. provided expertise in interpretation of bedform movement and sand flux in regards to Mars surface evolution and climate models. S. M. produced the digital elevation model. All authors shared ideas and results and helped produce the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.T.B. (nathan.bridges@jhuapl.edu).

doi:10.1038/nature11025

Resolving the time when an electron exits a tunnelling barrier

Dror Shafir[1]*, Hadas Soifer[1]*, Barry D. Bruner[1], Michal Dagan[1], Yann Mairesse[2], Serguei Patchkovskii[3], Misha Yu. Ivanov[4,5], Olga Smirnova[5] & Nirit Dudovich[1]

The tunnelling of a particle through a barrier is one of the most fundamental and ubiquitous quantum processes. When induced by an intense laser field, electron tunnelling from atoms and molecules initiates a broad range of phenomena such as the generation of attosecond pulses[1], laser-induced electron diffraction[2,3] and holography[2,4]. These processes evolve on the attosecond timescale (1 attosecond ≡ 1 as = 10^{-18} seconds) and are well suited to the investigation of a general issue much debated since the early days of quantum mechanics[5–7]—the link between the tunnelling of an electron through a barrier and its dynamics outside the barrier. Previous experiments have measured tunnelling rates with attosecond time resolution[8] and tunnelling delay times[9]. Here we study laser-induced tunnelling by using a weak probe field to steer the tunnelled electron in the lateral direction and then monitor the effect on the attosecond light bursts emitted when the liberated electron re-encounters the parent ion[10]. We show that this approach allows us to measure the time at which the electron exits from the tunnelling barrier. We demonstrate the high sensitivity of the measurement by detecting subtle delays in ionization times from two orbitals of a carbon dioxide molecule. Measurement of the tunnelling process is essential for all attosecond experiments where strong-field ionization initiates ultrafast dynamics[10]. Our approach provides a general tool for time-resolving multi-electron rearrangements in atoms and molecules[11–13]—one of the key challenges in ultrafast science.

Strong-field light–matter interactions present a unique combination of quantum and semiclassical physics. A broad range of strong-field phenomena is initiated when a strong laser field suppresses the atomic or molecular Coulomb barrier and thereby induces tunnel ionization[14,15]. Although tunnelling is a purely quantum phenomenon, once the electron is freed its motion in the strong laser field becomes nearly classical and can be described in terms of trajectories[16,17]. The liberated electron oscillates in the laser field, and can return to recombine with the parent ion. Using the radiation emitted during this process, which is known as high-harmonic generation (ref. 10 and references therein) we observe substantial deviations from the semi-classical model[9,16,18] that assumes that the instantaneous velocity of the electron emerging from the barrier is equal to zero.

High-harmonic generation offers ångström-scale spatial resolution of the electronic structure of the parent ion, as a result of the de Broglie wavelength of the electron, and ångström-scale spatial resolution of the continuum electron motion, as a result of the size of the ion in the ground state with which the continuum electron must recombine. High-harmonic generation also offers attosecond temporal resolution, owing to the connection between the ionization time (t_i), the return time (t_r) and the return energy ($E_r(t_r)$) of the corresponding electron trajectory[19] mapped onto the energy of the emitted photon (Ω) by the expression of energy conservation $E_r(t_r) + I_p = \Omega$ (where I_p is the ionization potential and atomic units are used throughout) as described in Fig. 1. Although the link between t_r and E_r has been studied using several different schemes[10,20–22], the subcycle mapping between t_i and E_r has until now been hidden in strong-field experiments.

The independent characterization of ionization and recombination times requires a probe that is both perturbative and fast enough to monitor these electron trajectories on the subcycle timescale. Such a probe is introduced by adding a weak second-harmonic field, $\mathbf{F}_{2\omega} = F_{2\omega}\cos(2\omega t + \phi)\mathbf{e}_y$, to the strong fundamental beam, $F_{\omega}\cos(\omega t)\mathbf{e}_x$. Here ω is the fundamental frequency, ϕ is the controlled delay between the two fields (colours, corresponding to frequencies ω and 2ω) and the fields are orthogonally polarized ($\mathbf{e}_y$ and $\mathbf{e}_x$ are unit vectors in the y and x directions, respectively).

In the tunnel ionization regime[14], the weak probe field mostly affects the electron after tunnelling and thereby signals its exit from the barrier (Supplementary Information). As the electron moves in the continuum, the weak probe field displaces it laterally, therefore suppressing recombination[23] (Fig. 2a). This displacement acts as a gate—for each delay ϕ, only some trajectories (Fig. 2a, red), launched in a narrow range of times, return to the parent ion and contribute to the emitted signal. As ϕ is changed, the 'displacement gate' shifts and a different ionization window is selected (Fig. 2b). By measuring the harmonic intensity as a function of ϕ, we can find the exit time (t_i) of the corresponding electron trajectory. The displacement depends on the full travel time, that is, both t_i and t_r. The return time (t_r) can be extracted by considering the lateral velocity component, v_y, of the returning electron. The probe field breaks the mirror symmetry of the electron motion in the two subsequent half-cycles of the fundamental field, which is reflected by the emission of even harmonics of the fundamental field[23]. The symmetry breaking is maximal when the velocity of the returning electron at the moment of recombination (t_r) has a

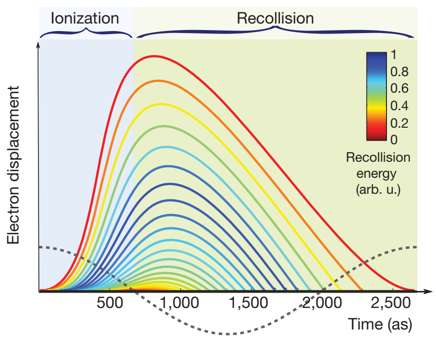


Figure 1 | Electron trajectories contributing to the recollision process. The coloured lines represent the spatio-temporal description of various trajectories; each colour encodes a recolliding energy, increasing from red to blue. The black dashed line shows the electric field along the cycle. arb. u., arbitrary units.

[1]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel. [2]Université Bordeaux-CEA-CNRS, CELIA, UMR5107, F-33400 Talence, France. [3]National Research Council of Canada, 100 Sussex Drive, Ottawa, Ontario K1A 0R6, Canada. [4]Department of Physics, Imperial College London, South Kensington Campus, London SW7 2AZ, UK. [5]Max-Born Institute for Nonlinear Optics and Short Pulse Spectroscopy, Max-Born-Strasse 2A, D-12489 Berlin, Germany.
*These authors contributed equally to this work.

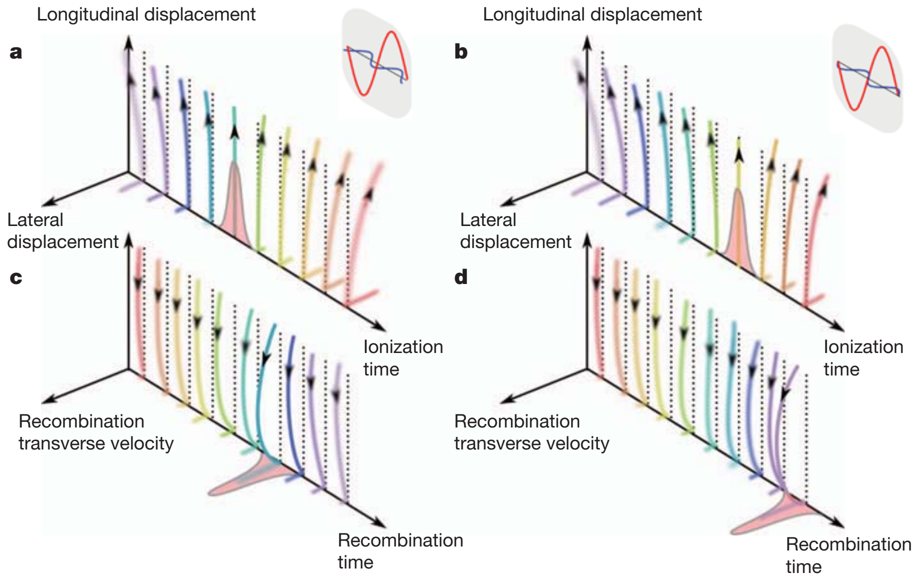


Figure 2 | Schematic description of the two-colour gates. a, **b**, Displacement gate. Different electron trajectories are separated along the ionization time axis (**a**). The displacement gate induces a lateral shift that suppresses the return probability. Only a narrow region in time is selected by the gate (shaded red). As the two-colour delay is changed, the gate is shifted within the optical cycle; a delayed ionization window is selected (**b**). **c**, **d**, Velocity gate. The different trajectories are separated in recombination time (**c**). The velocity gate controls the lateral velocity and, hence, the angle at recombination. A narrow region in time corresponds to the maximal recombination angle, selected by the gate (shaded red). Modifying the two-colour delay shifts the velocity gate within the optical cycle, thereby choosing a delayed recombination window (**d**).

maximal lateral component v_y (Fig. 2c, d). This component is directly mapped on the intensity ratio of the adjacent even and odd harmonics[23], which is maximized for maximal v_y. This ratio, measured as a function of ϕ, provides an additional independent observable ('velocity gate') that allows us to disentangle t_i and t_r.

In our experiment, we generate high harmonics from helium atoms using a strong fundamental field with a wavelength of $\lambda = 800$ nm and a weak second-harmonic field with an intensity 1% of that of the fundamental field (Supplementary Information). In Fig. 3a, we present the normalized harmonic intensity as a function of harmonic order and ϕ for laser intensity $I \approx 3.8 \times 10^{14}\,\mathrm{W\,cm^{-2}}$. For each harmonic, we extract the delay (ϕ^{y}_{max}) where the signal (sum of adjacent even and odd harmonics, see Supplementary Information) is maximized by the displacement gate. The delay ϕ^{y}_{max} changes with harmonic order, reflecting its dependence on the trajectory. In addition, we measure the ratio between adjacent even and odd harmonics, which is set by the velocity gate (Fig. 3b). Again, we extract the delay (ϕ^{v}_{max}) where this ratio is maximized by the velocity gate. The ϕ dependences of the two observables are significantly different, reflecting their independence.

Taking into account the perturbative nature of the probe field, we map the measured values of ϕ^{y}_{max} and ϕ^{v}_{max} onto t_i and t_r using the displacement and velocity gate equations (Supplementary Information). Calculating the trajectory shift introduced by the second colour, we take into account that all interactions except the probing field dictate t_i and t_r. Although the core potential modifies the unperturbed trajectory, it contributes only at higher order to the weak perturbation introduced by the second colour[21] (Supplementary Information). In our experiment, ϕ is not directly measured and is known only up to a constant phase shift, and such uncertainty leads to an absolute time shift in the reconstructed times. We therefore set ϕ by fixing the averaged recombination times to be the theoretical times taken from the quantum orbit model[24] (Supplementary Information).

In Fig. 3c, we present the reconstructed ionization and recombination times (red dots), and the predictions of the standard semiclassical model[16] (grey curves), which assumes that after tunnelling the electron appears at the ion position with zero initial velocity. The reconstructed recombination times agree well with the prediction of the semiclassical model (assuming the same harmonic cut-off). For the ionization times, a distinct deviation is observed for the lower harmonics, with the reconstructed ionization window about 100 as shorter than the semiclassical one. Including the offset of the trajectory from the ion after tunnelling does not alter this picture.

To understand the origin of this discrepancy, we return to the initial conditions set by the tunnelling process[24,25]. At the moment (t_0) when a bound electron enters the tunnelling barrier, its kinetic energy becomes negative[15]:

$$\frac{\mathbf{v}^2(t_0)}{2} = \frac{(\mathbf{p}+\mathbf{A}(t_0))^2}{2} = -I_p$$

Here I_p is the ionization potential, $\mathbf{v}(t) = \mathbf{p} + \mathbf{A}(t)$ is the instantaneous electron velocity, $\mathbf{p}$ is the electron drift (canonical) momentum and $\mathbf{A}(t)$ is the vector potential of the laser field. Hence, both $\mathbf{v}(t_0) = \mathbf{p} + \mathbf{A}(t_0)$ and t_0 are complex, the hallmarks of tunnelling. The quantum exit (ionization) time (t_i) is defined as the real part of the complex-valued time $t_0 = t_i + i\tau$ (ref. 15). For our experimental conditions, it differs from the semiclassical exit time by up to a few hundred attoseconds (compare the grey curve (semiclassical) and the black curve (quantum) in Fig. 3c). The reconstructed ionization times (red dots) agree much better with this definition; the green curve illustrates the minor corrections to the reconstructed ionization times due to the effect of the weak probe on tunnelling (Supplementary Information). The difference between the quantum and semiclassical pictures of tunnelling reflects the non-zero velocity of the electron as it exits the barrier at t_i.

Our approach opens the way to observing tunnelling in more complex systems. Whereas in helium only one orbital participates in strong-field ionization, in many molecular systems several orbitals (channels) may contribute[12,26]. Owing to the proximity between their I_p values, only subtle differences in the ionization times and initial conditions are expected. Our approach can distinguish experimentally between ionization dynamics associated with different channels, as we now investigate.

In the absence of the weak gating field, $\mathbf{F}_{2\omega}$, and with two orbitals, labelled $i = 1, 2$, participating in ionization, the intensity of the Nth harmonic is given by a coherent superposition of the contributions from the two channels:

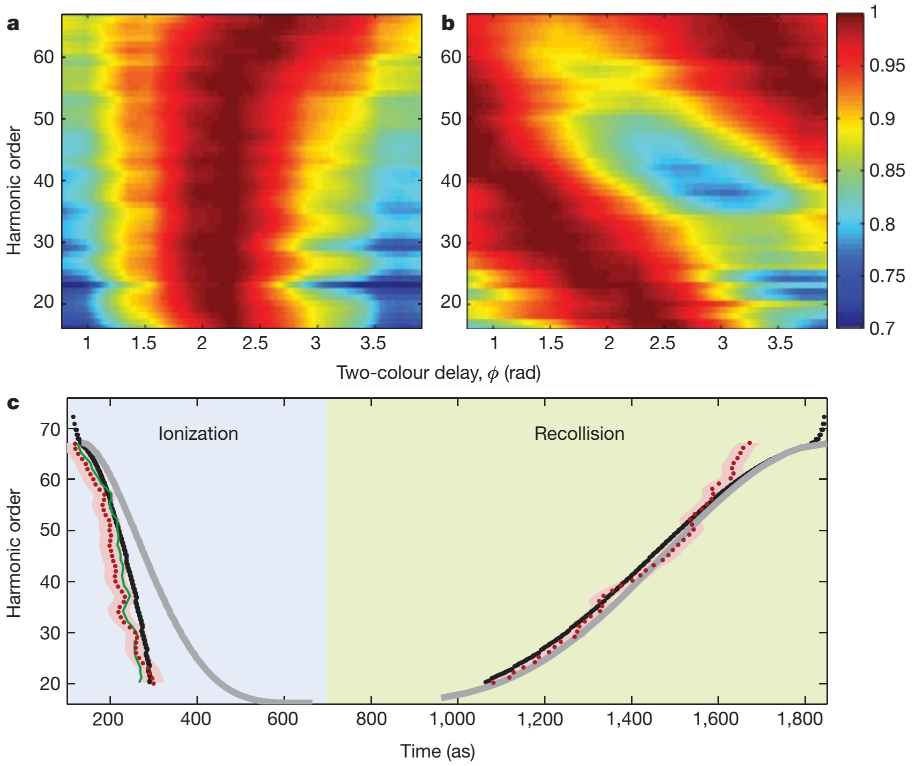


Figure 3 | Reconstruction of the ionization and recollision times. **a**, Displacement gate: normalized harmonic signal (colour scale) as a function of harmonic order and two-colour delay (ϕ) for helium. (For each harmonic order the signal is divided by the maximal signal for that harmonic. Colour scale shows the relative strength of the normalized signal.) **b**, Velocity gate: normalized recollision angle (colour scale) as a function of harmonic order and ϕ. (For each harmonic order the angle is divided by the maximal angle for that harmonic. Colour scale shows the relative strength of the normalized signal.) **c**, Reconstructed ionization and recollision times extracted from **a** and **b** (red dots). The pink shaded areas represent the uncertainty in the reconstruction procedure. The extracted times are compared to the calculated times according to the semiclassical model (grey curves) and the quantum stationary solution (black curves). The reconstructed ionization times using the combined ionization–displacement gate (green curve) are also shown (Supplementary Information).

$$S_N = S_N^{(1)} + S_N^{(2)} + 2\sqrt{S_N^{(1)} S_N^{(2)}}\cos(\Delta\varphi_N)$$

Here $S_N^{(i)}$ is the harmonic intensity for channel i and $\Delta\varphi_N$ is the relative phase between the channels, which can change by several π across the harmonic spectrum[12]. When the weak field $\mathbf{F}_{2\omega}$ is added, the signal associated with each channel is multiplied by the corresponding gate: $\tilde{S}_N^{(i)}(\phi) = S_N^{(i)}[G_N^{(i)}(\phi)]^2$, $i = 1, 2$. Here $G_N^{(i)}(\phi)$ is the two-colour (displacement) gate for the harmonic amplitudes.

For channels with close values of I_p, the values of $G_N^{(i)}(\phi)$ are also very similar, and it is possible to linearize the total signal as (Supplementary Information)

$$\tilde{S}_N(\phi) \approx G_N^2(\phi) S_N \left(1 + \frac{\Delta G_N(\phi)}{G_N(\phi)}\frac{\Delta S_N}{S_N}\right) \qquad (1)$$

where $G_N = (G_N^{(1)} + G_N^{(2)})/2$, $\Delta G_N = G_N^{(1)} - G_N^{(2)}$ represents the differential gate and $\Delta S_N = S_N^{(1)} - S_N^{(2)}$.

The second term in equation (1) describes the sensitivity of the measurement to small differences in ionization between the two channels, encoded in ΔG_N. In general, $\Delta G_N \ll G_N$ and $\Delta S_N \ll S_N$, and such detection therefore requires an extremely high signal-to-noise ratio. However, when the two channels destructively interfere, leading to dynamical minima in the harmonic spectra[12], $S_N^{(1)} \approx S_N^{(2)}$, $\cos(\Delta\varphi_N) = -1$ and we find that $\Delta S/S = (\sqrt{S^{(1)}} + \sqrt{S^{(2)}})/(\sqrt{S^{(1)}} - \sqrt{S^{(2)}}) \gg 1$. The second term in equation (1) then dominates and the measurement directly resolves the differential gate. As confirmed in our atomic experiments, for each channel the modulation of the signal ($G_N^{(i)}(\phi)$) follows a $\cos(2\phi)$ function, and so does G_N.

The differential gate (ΔG_N) is shifted by ~π/2 relative to the single-channel gate ($G_N^{(i)}(\phi)$) (Supplementary Information), leading to a different optimal phase ϕ_{max}^y in the gated signal. The observation of this phase shift would signify our ability to detect the differences in the ionization dynamics associated with each channel.

We demonstrate this ability using aligned carbon dioxide molecules. For molecules aligned parallel to the polarization of the ionizing field, two orbitals—HOMO (the highest occupied molecular orbital) and HOMO−2 (two orbitals below HOMO)—participate in ionization[12]. At around harmonic 29, we observe a spectral minimum arising from destructive interference between the contributions of the two orbitals[12]. The effect of differential gating is illustrated in Fig. 4a–c. Figure 4a shows the calculated quantum ionization times for the two carbon dioxide orbitals. Figure 4b schematically shows the modulation of the harmonics as a function of ϕ for each channel, emphasizing the subtle phase difference between the two signals. Figure 4c shows the total signal for constructive (black curve) and destructive (blue curve) interference of the two channels, demonstrating the expected phase shift. This phase shift is immediately visible in Fig. 4d, which shows a normalized harmonic spectrum calculated for the carbon dioxide molecule using the multichannel method described in ref. 12, with ionization rates from ref. 27. The destructive interference of the two channels at harmonics 27–29 is marked by a substantial shift of the optimal two-colour delay by ~2 rad.

Figure 4e, f presents the normalized experimental spectrum as a function of ϕ, at $I \approx 1.3 \times 10^{14}\,\mathrm{W\,cm^{-2}}$ and for a second-harmonic field with an intensity 2% of that of the fundamental field. For perpendicular alignment (Fig. 4e), a single orbital (HOMO) dominates the signal[12] and ϕ_{max}^y changes smoothly with the harmonic order. For

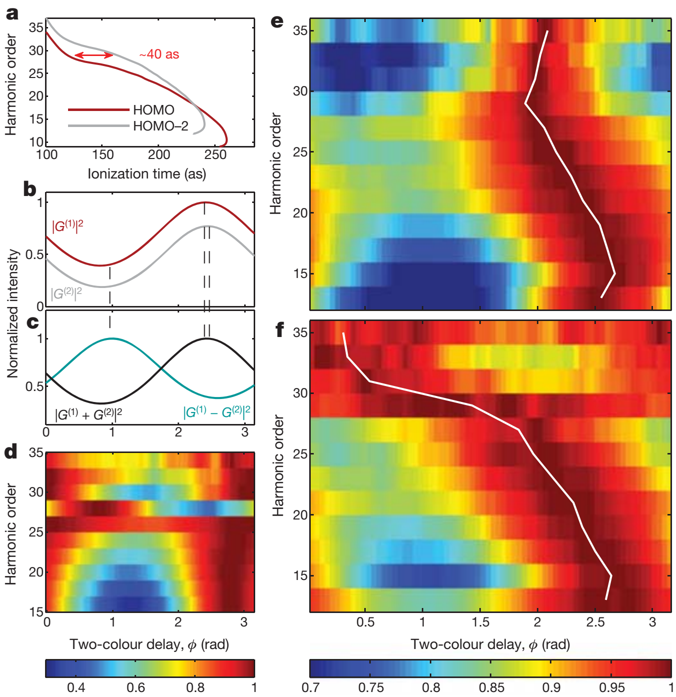


Figure 4 | Gating two-channel tunnel ionization in aligned carbon dioxide molecules. a, Two stationary solutions associated with the HOMO (red curve) and HOMO−2 (grey curve) orbitals, calculated for $I = 1.3 \times 10^{14}$ W cm^{-2}. **b**, Calculation of $|G^{(1)}|^2$ and $|G^{(2)}|^2$ as functions of the two-colour delay (ϕ). The black dashed lines indicate ϕ^{y}_{max} for each channel. **c**, Calculation of G_N for constructive (black curve) and destructive (blue curve) interference. **d**, Theoretical calculation of the normalized harmonic signal for a carbon dioxide molecule aligned at an angle of $\pm 35^\circ$ to the 40-fs, 800-nm, $I = 1.3 \times 10^{14}$ W cm^{-2} fundamental field, using the second-harmonic field at a 2% intensity level. **e**, **f**, Normalized measured harmonic spectra (colour scale) as functions of harmonic order and ϕ for carbon dioxide molecules aligned at 90° (**e**) and 0° (**f**). The white curves follow the phases (ϕ^{y}_{max}) that maximize the harmonic signal. The two colour phase was shifted by 0.3 rad in panels **d**–**f** to show the full extent of the phase shift.

parallel molecular alignment, where two orbitals contribute to the signal (Fig. 4f), ϕ^{y}_{max} shifts by $\sim\pi/2$ at around harmonic 29 (corresponding to the dynamical minimum[12]) and the overall shift is about 2 rad. This drastic effect agrees with the above theoretical analysis. For a lower laser intensity, the phase jump moves to a lower harmonic, corresponding to the new position of the dynamical minimum (not shown).

In many attosecond time-resolved experiments, tunnel ionization serves as a pump, thus starting the clock, and the recollision serves as a probe[12,19,28,29]. Our experiment, which links the ionization time (t_i) to the return time (t_r), calibrates the internal attosecond clock on which the experiments are based.

In most molecular systems, ionization involves attosecond core rearrangements[13,30], which may lead to a real time delay associated with non-trivial phases between different tunnelling channels[30]. In such cases, a direct *in situ* measurement, such as that presented here, becomes indispensable. Time-resolving the tunnelling dynamics in complex molecular systems will provide deep insights into fundamental multi-electron phenomena, which is the long-term goal of attosecond science.

Received 6 September 2011; accepted 2 March 2012.

1. Hentschel, M. *et al.* Attosecond metrology. *Nature* **414,** 509–513 (2001).
2. Spanner, M., Smirnova, O., Corkum, P. B., Ivanov, M. & Yu.. Reading diffraction images in strong field ionization of diatomic molecules. *J. Phys. B* **37,** L243–L250 (2004).
3. Meckel, M. *et al.* Laser-induced electron tunneling and diffraction. *Science* **320,** 1478–1482 (2008).
4. Huismans, Y. *et al.* Time-resolved holography with photoelectron waves. *Science* **331,** 61–64 (2011).
5. Büttiker, M. & Landauer, R. Traversal time for tunneling. *Phys. Rev. Lett.* **49,** 1739–1742 (1982).
6. Steinberg, M. A., Kwiat, P. G. & Chiao, R. Y. Measurement of the single-photon tunnelling time. *Phys. Rev. Lett.* **71,** 708–711 (1993).
7. Landauer, R. & Martin, Th Barrier interaction time in tunneling. *Rev. Mod. Phys.* **66,** 217–228 (1994).
8. Uiberacker, M. *et al.* Attosecond real-time observation of electron tunnelling in atoms. *Nature* **446,** 627–632 (2007).
9. Eckle, P. *et al.* Attosecond ionization and tunneling delay time measurements in helium. *Science* **322,** 1525–1529 (2008).
10. Krausz, F., Ivanov, M. & Yu.. Attosecond physics. *Rev. Mod. Phys.* **81,** 163–234 (2009).
11. Schultze, M. *et al.* Delay in photoemission. *Science* **328,** 1658–1662 (2010).
12. Smirnova, O. *et al.* High harmonic interferometry of multi-electron dynamics in molecules. *Nature* **460,** 972–977 (2009).
13. Walters, Z. B. & Smirnova, O. Attosecond correlation dynamics during electron tunnelling from molecules. *J. Phys. B* **43,** 161002 (2010).
14. Keldysh, L. V. Ionization in the field of a strong electromagnetic wave. *Sov. Phys. JETP* **20,** 1307–1314 (1965).
15. Perelomov, M. A., Popov, S. V. & Terent'ev, M. V. Ionization of atoms in an alternating electric field: II. *Sov. Phys. JETP* **24,** 207 (1967).
16. Corkum, P. B. Plasma perspective on strong field multiphoton ionization. *Phys. Rev. Lett.* **71,** 1994–1997 (1993).
17. Salières, P. *et al.* Feynman's path-integral approach for intense-laser-atom interactions. *Science* **292,** 902–905 (2001).
18. Pfeiffer, A. N. *et al.* Attoclock reveals natural coordinates of the laser-induced tunnelling current flow in atoms. *Nature Phys.* **8,** 76–80 (2012).
19. Baker, S. *et al.* Probing proton dynamics in molecules on an attosecond time scale. *Science* **312,** 424–427 (2006).
20. Mairesse, Y. *et al.* Attosecond synchronization of high-harmonic soft X-rays. *Science* **302,** 1540–1543 (2003).
21. Dudovich, N. *et al.* Measuring and controlling the birth of attosecond XUV pulses. *Nature Phys.* **2,** 781–786 (2006).
22. Chirilă, C. C., Dreissigacker, I., van der Zwan, E. V. & Lein, M. Emission times in high-order harmonic generation. *Phys. Rev. A* **81,** 033412 (2010).
23. Shafir, D. *et al.* Atomic wavefunctions probed through strong-field light-matter interaction. *Nature Phys.* **5,** 412–416 (2009).
24. Lewenstein, M., Balcou, Ph, Ivanov, M., Yu., L'Huillier, A. & Corkum, P. B. Theory of high-harmonic generation by low-frequency laser fields. *Phys. Rev. A* **49,** 2117–2132 (1994).
25. Dahlström, J. M., L'Huillier, A. & Mauritsson, J. Quantum mechanical approach to probing the birth of attosecond pulses using a two-colour field. *J. Phys. B* **44,** 095602 (2011).
26. Haessler, S. *et al.* Attosecond imaging of molecular electronic wavepackets. *Nature Phys.* **6,** 200–206 (2010).
27. Murray, R., Spanner, M., Patchkovskii, S., Ivanov, M. & Yu.. Tunnel ionization of molecules and orbital imaging. *Phys. Rev. Lett.* **106,** 173001 (2011).
28. Tong, X. M., Zhao, Z. X. & Lin, C. D. Probing molecular dynamics at attosecond resolution with femtosecond laser pulses. *Phys. Rev. Lett.* **91,** 233203 (2003).
29. Niikura, H. *et al.* Probing molecular dynamics with attosecond resolution using correlated wave packet pairs. *Nature* **421,** 826–829 (2003).
30. Mairesse, Y. *et al.* High harmonic spectroscopy of multichannel dynamics in strong-field ionization. *Phys. Rev. Lett.* **104,** 213601 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements N.D. acknowledges the Minerva Foundation, the Israeli Science Foundation and the Crown Center of Photonics for financial support. N.D. and O.S. acknowledge the German-Israeli Foundation, and D.S. acknowledges support from the Converging Technologies Fellowship of the Israeli Ministry of Science. H.S. is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities. O.S. acknowledges the support of the DFG grant Sm292/2-1, and M.Yu.I. acknowledges the support of the EPSRC programme grant EP/I032517/1. Y.M. acknowledges financial support from the ANR (ANR-08-JCJC-0029 HarModyn) and the Conseil Regional d'Aquitaine (20091304003 AttoMol). M.I. acknowledges support of the EPSRC Programme Grant EP/I0325171/1.

Author Contributions D.S., H.S., B.D.B., M.D. and N.D. designed, performed and analysed the experiment. Y.M. contributed to the analysis. M.Yu.I. and O.S. developed the theory and performed the calculations of harmonic spectra. S.P. provided the quantum chemistry input for the calculations. All authors contributed to writing the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.D. (nirit.dudovich@weizmann.ac.il).

doi:10.1038/nature11122

Light-induced liquid crystallinity

Tamas Kosa[1]*, Ludmila Sukhomlinova[1], Linli Su[1], Bahman Taheri[1], Timothy J. White[2]* & Timothy J. Bunning[2]

Liquid crystals are traditionally classified as thermotropic, lyotropic or polymeric, based on the stimulus that governs the organization and order of the molecular system[1]. The most widely known and applied class of liquid crystals are a subset of thermotropic liquid crystals known as calamitic, in which adding heat can result in phase transitions from or into the nematic, cholesteric and smectic mesophases. Photoresponsive liquid-crystal materials and mixtures can undergo isothermal phase transitions if light affects the order parameter of the system within a mesophase sufficiently. In nearly all previous examinations, light exposure of photoresponsive liquid-crystal materials and mixtures resulted in order-decreasing photo-induced isothermal phase transitions[2]. Under specialized conditions, an increase in order with light exposure has been reported, despite the tendency of the photoresponsive liquid-crystal system to reduce order in the exposed state[3–7]. A direct, photo-induced transition from the isotropic to the nematic phase has been observed in a mixture of spiropyran molecules and a nematic liquid crystal[8]. Here we report a class of naphthopyran-based materials that exhibit photo-induced conformational changes in molecular structure capable of yielding order-increasing phase transitions. Appropriate functionalization of the naphthopyran molecules leads to an exceedingly large order parameter in the open form, which results in a clear to strongly absorbing dichroic state. The increase in order with light exposure has profound implications in optics, photonics, lasing and displays and will merit further consideration for applications in solar energy harvesting. The large, photo-induced dichroism exhibited by the material system has been long sought in ophthalmic applications such as photochromic and polarized variable transmission sunglasses.

Naphthopyran was targeted as the photochromic unit of the molecules examined here because of the ease with which the absorbance in the open form can be tuned, the relative simplicity of the chemistry, and the colourless degradation products that are important in commercial ophthalmic applications as variable transmission lenses. This work primarily examines the naphthopyran molecule methyl 8-(4′-pentylbiphenyl-4-yl)-2-phenyl-2-(4-fluorophenyl)-2H-naphtho[1,2-b]pyran-5-carboxylate (**1** in Fig. 1a), although other naphthopyrans were synthesized (Supplementary Fig. 2 and Supplementary Table 1). The chemical structure and the photo-induced conformational change of molecule **1** from the closed form to the open form are illustrated in Fig. 1a. Molecule **1** becomes elongated and planar in the open form. The photochromism of **1**, when mixed at 1 wt% into the nematic liquid crystal mixture ZLI-4788, is shown in Fig. 1b. The absorbance spectra of **1** undergoes a bathochromic shift from the ultraviolet to the visible spectrum on irradiation with light at a wavelength of 365 or 405 nm. Importantly for practical utility, the photo-induced changes are short-lived and reversible, as the material biexponentially decays back to closed form in a few minutes (Fig. 1b inset)[9]. Because of the mesogenic substitution of the *n*-pentyl biphenyl onto the naphthopyran core, **1** exhibits excellent solubility in liquid-crystal hosts. Owing to the shape of the closed form of **1**, doping increasing concentrations of this molecule into a model, single-component liquid-crystal host such as *n*-pentyl cyanobiphenyl (5CB) shifts the nematic to isotropic phase

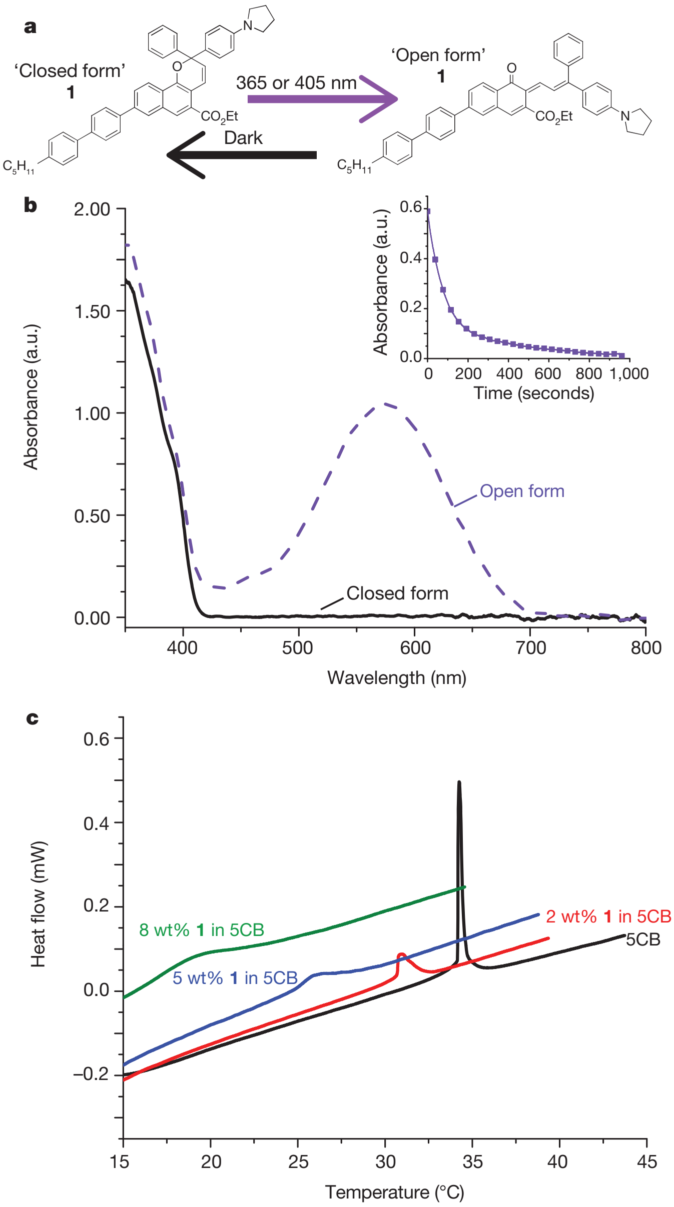


Figure 1 | Naphthopyran material properties. **a**, Photo-induced ring opening of methyl 8-(4′-pentylbiphenyl-4-yl)-2-phenyl-2-(4-fluorophenyl)-2H-naphtho[1,2-b]pyran-5-carboxylate (**1**) upon exposure to ultraviolet light of wavelength 365 or 405 nm. **b**, Bathochromic shift in absorbance of **1** (1 wt% in ZLI-4788) on exposure to 405 nm irradiation at 2 mW cm^{-2} (solid line, closed form; dashed line, open form). a.u., arbitrary units. Inset, the biexponential decay of **1** (measured by absorbance at 600 nm) from the open form to the closed form is rapid, being essentially complete after ~1,000 s. **c**, DSC thermograms of mixtures of **1** with the room-temperature nematic liquid crystal *n*-pentyl cyanobiphenyl (5CB). Increasing the concentration of **1** reduced the nematic-to-isotropic transition temperature of the **1**/5CB mixtures, confirming that **1** is order-disrupting in the closed form.

[1]Alpha Micron Inc., Kent, Ohio 44240, USA. [2]Air Force Research Laboratory, Materials and Manufacturing Directorate, Wright Patterson Air Force Base, Ohio 45433, USA.
*These authors contributed equally to this work.

transition temperature (an exothermic peak in differential scanning calorimetry, DSC) to lower temperatures (Fig. 1c). This shift confirms that the closed form (ground state) of **1** is order-disrupting when mixed as a guest molecule into liquid-crystal hosts.

The capability of **1** to induce disordered-to-ordered and order-increasing phase transitions was examined in two representative systems by observing the materials using polarized optical microscopy (POM) before, during and after exposure to ultraviolet light (5 mW cm^{-2}, 365 nm wavelength). POM is a technique often used to characterize liquid-crystal phase transitions, which appear as characteristic textures associated with the birefringence of nematic, smectic and cholesteric liquid-crystal phases. The mixture **1** (2.4 wt%)/5CB was heated above the isotropic phase transition temperature (31.8 °C; Fig. 2a). Owing to the lack of birefringence in the isotropic state of the mixture, the image (Fig. 2a) is black. As is evident in Fig. 2b, exposure to ultraviolet light generated a birefringent texture in the cell, typical of the nematic phase. This photo-induced nematic phase returned to the isotropic state (Fig. 2c) within five seconds after removal of the ultraviolet light. Order-increasing phase transitions between liquid-crystal phases can also be induced, demonstrated here in a mixture of **1** (1.8 wt%) and *n*-octyl cyanobiphenyl (8CB). 8CB is a single component liquid crystal with a room temperature smectic A phase and exhibits a smectic A to nematic transition at 34 °C and a nematic to isotropic transition at 41 °C. With the addition of 1.8 wt% **1**, the phase transition temperature for smectic A to nematic in 8CB decreased to 30.5 °C while the nematic to isotropic phase transition temperature decreased to 39.0 °C. As shown in Fig. 2d, the POM micrograph at 32.0 °C exhibited a texture characteristic of a nematic phase. Ultraviolet exposure of this **1**/8CB mixture at 32.0 °C induced a transition from the nematic to the smectic A phase. Once again, on removal of the ultraviolet light the material returned to the nematic phase nearly immediately, confirming the phototropic nature of the phase. After allowing **1** to revert to the closed form, the sample was heated to 40.3 °C to induce a transition of the **1**/8CB mixture into the isotropic phase. As demonstrated in Fig. 2g–i, ultraviolet exposure can also induce a phase transition from the isotropic to nematic phase in the **1**/8CB mixture. Importantly for applications, this material system rapidly returns to the original mesophase on removal of irradiation. Identical phase transitions are also observable in aligned cells (Supplementary Fig. 3).

Building on the ability to induce the full gamut of phase transitions, we demonstrate the potential utility of light-induced liquid crystallinity in the cholesteric liquid-crystal phase and the twisted nematic geometry. A cholesteric liquid-crystal mixture capable of photo-induced phase transitions was formulated by mixing the commercially available chiral dopant R1011 with 5CB. To this mixture, 4.0 wt% of **1** was added, disrupting the cholesteric phase and forming an isotropic liquid at room temperature (Fig. 3A, a). As shown in the POM images in Fig. 3A, ultraviolet light induced an isotropic to cholesteric phase transition, evident in the Grandjean texture in Fig. 3A, b. Removal of light returned the **1**/R1011/5CB mixture to the isotropic state. The optical properties of the photo-induced cholesteric phase are further elucidated in Fig. 3B, which shows a plot of the transmission spectra of the mixture before and during ultraviolet exposure. As evident in Fig. 3B, the mixture was transparent in the visible portion of the electromagnetic spectrum before light exposure (no absorption or reflection). However, on exposure to 365 nm light an absorption band appeared at 603 nm while the reflection bandgap from the cholesteric phase appeared at 800 nm. Although intentionally separated in Fig. 3B, the absorption and reflection bands can easily be overlapped to yield high-contrast, dichroic, optical materials. Spatial patterning of the photoresponsive material system is demonstrated in Fig. 3C, which illustrates the optical response of a twisted nematic cell composed of the mixture **1**/5CB (see also Supplementary Movie 1). The cell was illuminated with polarized white light from below and viewed through a crossed analyser. Accordingly, the cell initially appeared dark, as the mixture is isotropic at room temperature. On exposure to 405 nm light through the mask, the exposed areas undergo a transition into the nematic phase, permitting light to pass through the crossed analyser. On removal of the irradiation, the photo-induced nematic phase was retained only briefly.

In addition to enabling the distinctive ability to use light to induce liquid crystallinity, the open form of **1** can exhibit a large order parameter (0.722) that greatly exceeds prior reports[10,11] and rivals the order

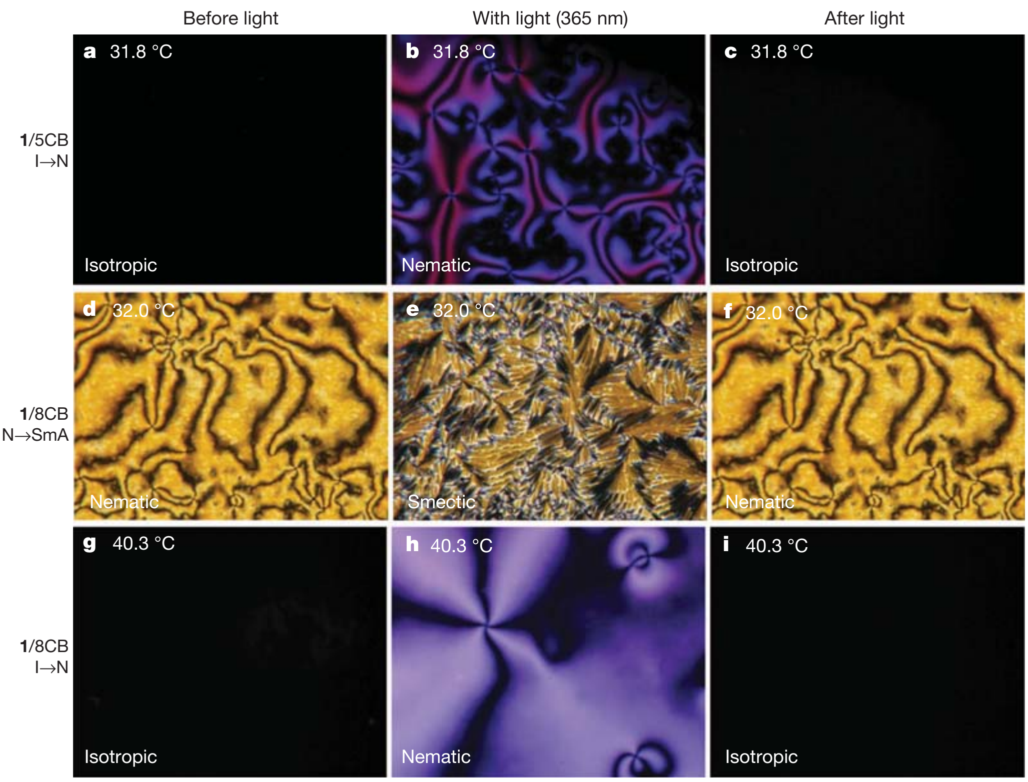


Figure 2 | Increasing order. Polarized optical micrographs of phototropic phase transitions of **1** (2.4 wt%)/5CB and **1** (1.8 wt%)/8CB. Images were taken before light exposure, during light exposure (365 nm ultraviolet light, 5 mW cm^{-2}), and after light exposure. Panels **a**–**c** illustrate the phototropic isotropic to nematic transition observed in the mixture **1**/5CB, at 31.8 °C. Panels **d**–**f** confirm that higher order transitions can be observed, such as a photo-induced nematic to smectic transition in the **1**/8CB mixture observed at 32.0 °C. The phototropic transition from isotropic to nematic is also observed in **1**/8CB at 40.3 °C, as shown in **g**–**i**. Phase transitions in aligned cells are included in Supplementary Fig. 3.

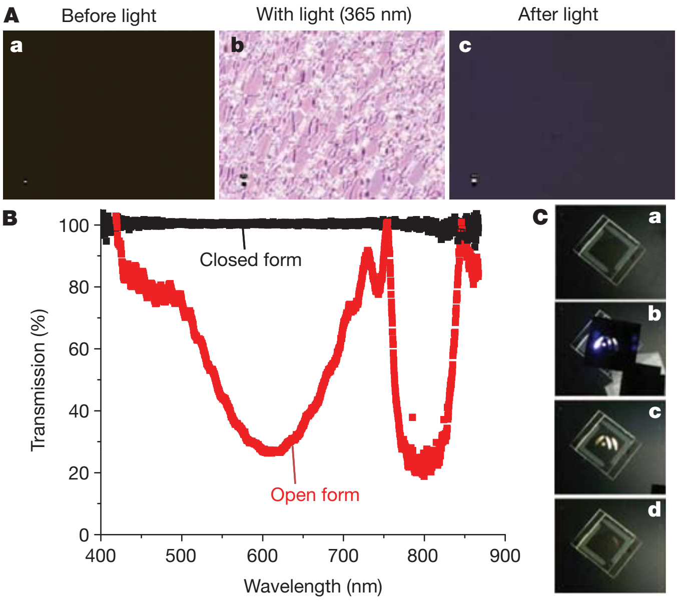


Figure 3 | Clear to coloured. Phototropic phase transition of **1** (4.0 wt%) mixed with R1011 (6 wt%)/5CB. **A**, Polarized optical micrographs of phototropic phase transitions with exposure to 365 nm ultraviolet light at 5 mW cm^{-2}: **a**, before exposure; **b**, during exposure; and **c**, after exposure. **B**, Transmission spectra of mixture before (black squares) and during (red squares) ultraviolet exposure. **C**, Spatial patterning with a mask is demonstrated in a twisted nematic cell placed on a polarized light source viewed through a crossed analyser: **a**, before exposure; **b**, during exposure. On removal of the light, the light is transmitted through the exposed area (**c**) before quickly vanishing (**d**). See also Supplementary Movie 1.

parameter of standard dichroic dyes. The order parameter and chemical structures of **1** are contrasted with two related naphthopyran molecules in Supplementary Fig. 2 and Supplementary Table 1, to illustrate the importance of mesogenic substitution. The photo-induced dichroism achievable in the isotropic to nematic transition of the **1**/ZLI-4788 mixture is presented in the absorption spectra of Fig. 4A. During continuous exposure to 405 nm irradiation, the dye exhibited considerable absorption when the white light probe was polarized parallel to the nematic director of the alignment cell. When the polarization of the white light probe was orthogonal to the nematic director of the alignment cell, the material exhibited limited absorption. The photo-induced dichroism is further illustrated in Fig. 4B, a–c. Before exposure, the mixture exhibited no dichroism (Fig. 4B, a). During and immediately after irradiation with 405 nm light, the mixture becomes strongly dichroic, as depicted in Fig. 4B, b and c.

In summary, light-induced liquid crystallinity has been demonstrated in a material system in which light exposure directly increased the order of the mixture through conformational changes resulting in mesophase transitions from lower to higher order. The order-increasing phase transitions realized here have been enabled by the synthesis of a new class of naphthopyran dyes with an order-disrupting closed form and order-enhancing open form. The dramatic increase in the order of the dye with light exposure is the driving mechanism of the observed behaviour. In addition to realizing order-increasing phase transitions, the naphthopyran molecules employed here exhibit unprecedented changes in the order parameter of the dye that rival those observed in standard dichroic dyes. Accordingly, photo-induced isotropic to nematic transition switch the transmission through the cell from clear to strongly absorbing and dichroic. The clear to dichroic transition has been long-pursued in ophthalmic applications, such as polarized variable transmission lenses. Future work aims to examine new photochromic molecules, begin systematic development of material formulations with wider temperature ranges for the photo-induced mesophases, and demonstrate potential opportunities afforded by this new type of photo-induced liquid-crystal phase change.

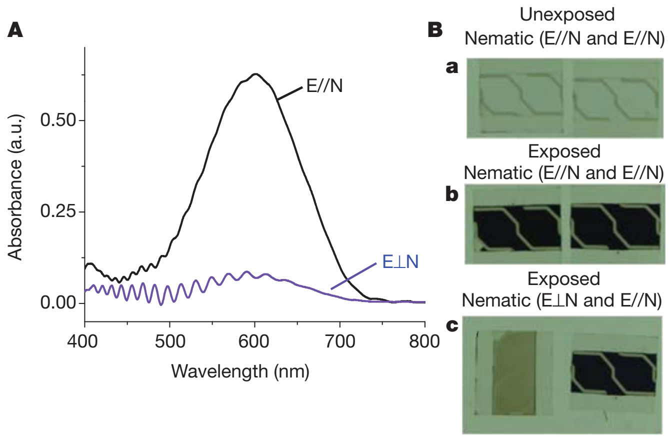


Figure 4 | Polarizing the light. **A**, Photo-induced dichroism observable in a mixture of **1** (1 wt%)/ ZLI-4788 subjected to a continuous exposure of 405 nm irradiation at 2 mW cm^{-2}. When the polarization (E) of the white light probe is parallel to the nematic director of the liquid-crystal mixture, the bathochromic absorbance of **1** in the open form (peak centred at 603 nm) is strong (black line). Rotating the polarization (E) of the white light probe such that it is orthogonal to the nematic director of the liquid-crystal mixture reveals the large, photo-induced dichroism of **1** (blue line). **B**, Images of two cells containing a mixture of **1** (1 wt%)/ZLI-4788 on a polarized light source. Before irradiation with 405 nm light, the cells are optically clear (**a**). On irradiation with 405 nm light, the conformational change of **1** from the open form results in strong absorbance when the nematic director is aligned parallel to the polarization (E) of light emitted from the source (**b**). Rotating the left-most cell 90° visually illustrates the strong photo-induced dichroism realizable in these materials (**c**).

Received 18 November 2011; accepted 28 March 2012.

1. Collings, P. & Hird, M. *An Introduction to Liquid Crystals: Chemistry and Physics* (Taylor and Francis, 1997).
2. Ikeda, T. & Tsutsumi, O. Optical switching and image storage by means of azobenzene liquid-crystal films. *Science* **268,** 1873–1875 (1995).
3. Ortica, F. *et al.* Effects of the environment on the photochromic behaviour of a novel indeno-fused naphthopyran. *Photochem. Photobiol. Sci.* **1,** 803–808 (2002).
4. Prasad, S. K., Nair, G. G. & Hegde, G. Dynamic self-assembly of the liquid-crystalline smectic A phase. *Adv. Mater.* **17,** 2086–2091 (2005).
5. Prasad, S. K., Nair, G. G. & Hegde, G. Nonequilibrium liquid crystalline layered phase stabilized by light. *J. Phys. Chem. B* **111,** 345–350 (2007).
6. Serak, S. V., Tabiryan, N. V. & Bunning, T. J. Nonlinear transmission of photosensitive cholesteric liquid crystals due to spectral bandwidth auto-tuning or restoration. *J. Nonlinear Opt. Phys. Mater.* **16,** 471–483 (2007).
7. Zalar, B. *et al.* Deuteron NMR investigation of a photomechanical effect in a smectic-A liquid crystal. *Phys. Rev. E* **62** (2-A), 2252–2262 (2000).
8. Kurihara, S. *et al.* Isothermal phase transition of liquid crystals induced by photoisomerization of doped spiropyrans. *J. Chem. Soc. Faraday Trans.* **87,** 3251–3254 (1991).
9. Maafi, M. & Brown, R. G. Photophysics and kinetics of naphthopyran derivatives, part 2: Analysis of diarylnaphthopyran kinetics. Degeneracy of the kinetic solution. *Int. J. Chem. Kinet.* **37,** 717–727 (2005).
10. Frigoli, M. & Mehl, G. H. Room temperature photochromic liquid crystal [3H]-naphtho[2,1-b]pyrans-photochromism in the mesomorphic state. *Chem. Commun.* **18,** 2040–2041 (2004).
11. Shragina, L. *et al.* Searching for photochromic liquid crystals. Spironaphthoxazine substituted with a mesogenic group. *Liq. Cryst.* **7,** 643–655 (1990).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements T.J.W. and T.J.B. were supported by the US Air Force Office of Scientific Research and the Materials and Manufacturing Directorate of the US Air Force Office of Scientific Research. T.K., L. Sukhomlinova, L. Su and B.T acknowledge partial support through US Air Force contract FA8650-05-D-5807.

Author Contributions T.K., T.J.W., L. Sukhomlinova, L. Su, B.T. and T.J.B. designed experiments. L. Sukhomlinova conceived of and synthesized the molecules. T.K. and T.J.W. completed the experiments. T.J.W. prepared the figures and wrote the manuscript with assistance from T.K., B.T. and T.J.B.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to T.K. (tamas@alphamicron.com) or T.J.W. (timothy.white2@wpafb.af.mil).

doi:10.1038/nature11097

Recent Northern Hemisphere tropical expansion primarily driven by black carbon and tropospheric ozone

Robert J. Allen[1], Steven C. Sherwood[2], Joel R. Norris[3] & Charles S. Zender[4]

Observational analyses have shown the width of the tropical belt increasing in recent decades as the world has warmed[1]. This expansion is important because it is associated with shifts in large-scale atmospheric circulation[2–4] and major climate zones[5,6]. Although recent studies have attributed tropical expansion in the Southern Hemisphere to ozone depletion[7–10], the drivers of Northern Hemisphere expansion are not well known and the expansion has not so far been reproduced by climate models[11]. Here we use a climate model with detailed aerosol physics to show that increases in heterogeneous warming agents—including black carbon aerosols and tropospheric ozone—are noticeably better than greenhouse gases at driving expansion, and can account for the observed summertime maximum in tropical expansion. Mechanistically, atmospheric heating from black carbon and tropospheric ozone has occurred at the mid-latitudes, generating a poleward shift of the tropospheric jet[12], thereby relocating the main division between tropical and temperate air masses. Although we still underestimate tropical expansion, the true aerosol forcing is poorly known and could also be underestimated. Thus, although the insensitivity of models needs further investigation, black carbon and tropospheric ozone, both of which are strongly influenced by human activities, are the most likely causes of observed Northern Hemisphere tropical expansion.

Recent observational analyses show that the tropics have widened by 2°–5° latitude since 1979 (ref. 1). This evidence is based on several metrics, including a poleward shift of the Hadley cell[2], subtropical dry zones[5], and extratropical storm tracks[6]. A more recent estimate of the tropospheric jet shift[4], based on satellite-derived temperatures, suggests a smaller rate of expansion of 1.6°.

Tropical expansion occurs in model simulations forced by increasing greenhouse gases, thus suggesting a likely cause[1,13,14]. Model-predicted expansion rates, however, are significantly less than those observed[11]. This discrepancy may be related to the relatively short observational record, the large natural variability of some expansion metrics, or model deficiencies.

Several recent studies have focused on tropical expansion in the Southern Hemisphere, and the important contribution of stratospheric ozone depletion[7–10]. Less has been said about the causes of Northern Hemisphere expansion. Recent equilibrium simulations with atmospheric general circulation models have shown that direct heating of the troposphere, such as that caused by absorbing aerosols or tropospheric ozone, can drive expansion[15]. Although indirect aerosol effects may also be important and could yield the opposite response[16], they may be significantly overestimated in current general circulation models[17] and in any case they mainly cool the surface rather than the atmosphere.

Owing to increased combustion of fossil fuels and biofuels, black carbon aerosols have increased substantially over much of the Northern Hemisphere during the last few decades, particularly over southeast Asia, while decreasing over much of Europe (Fig. 1). Despite the geographically heterogeneous evolution, black carbon has

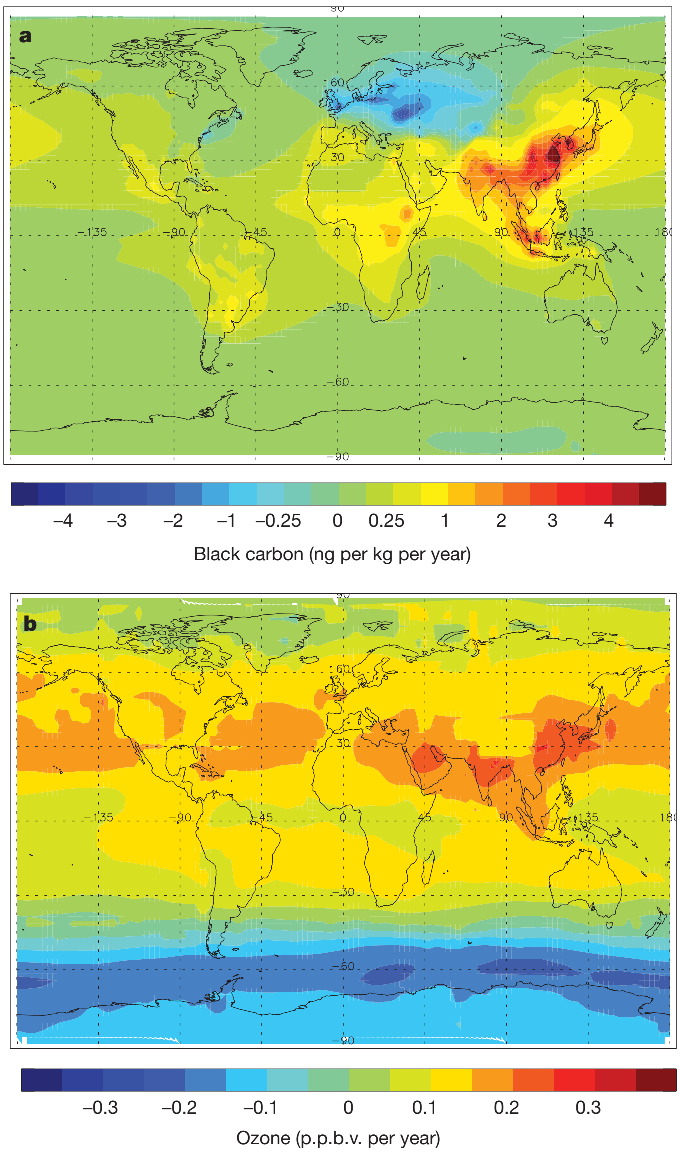


Figure 1 | 1970–2009 annual mean tropospheric trends. **a**, Black carbon; **b**, Ozone. p.p.b.v., parts per billion by volume. Black carbon concentration trends include hydrophobic and hydrophilic black carbon and are based on CAM simulations using CMIP5 black carbon emissions. Ozone trends come directly from the CMIP5 forcing data set.

[1]Department of Earth Sciences, University of California, Riverside 92521, USA. [2]Climate Change Research Centre and ARC Centre of Excellence for Climate Systems Science, University of New South Wales, Sydney 2052, Australia. [3]Scripps Institution of Oceanography, University of California, San Diego 92093, California, USA. [4]Earth System Science, University of California, Irvine 92697, California, USA.

increased monotonically since 1970 on average in the low and mid-latitudes, including the band 30°–50° N (Supplementary Figs 1–2), where recent studies show that heating can displace the tropical edge[12]. The same is true of tropospheric ozone, another indirect byproduct of combustion. Here, we investigate the transient effects of these atmospheric warming agents on Northern Hemisphere tropical width.

We quantify tropical width using a variety of metrics[5,11]: (1) the latitude of the tropospheric zonal wind maxima (JET); (2) the latitude where the Mean Meridional Circulation (MMC) at 500 hPa becomes zero on the poleward side of the subtropical maximum; (3) the latitude where precipitation minus evaporation (P − E) becomes zero on the poleward side of the subtropical minimum; (4) the latitude of the subtropical precipitation minimum (PMIN); and (5) the latitude of the subtropical cloud cover minimum over oceans (CMIN). To obtain an overall measure of tropical expansion, we also average the trends of all five metrics into a combined metric called 'ALL'. Expansion figures quoted in the text will be based on ALL unless otherwise specified.

Figure 2 compares the annual-mean poleward displacement of each metric, with 2σ uncertainty, between observations and a much-studied set of twentieth-century (Coupled Model Intercomparison Project version 3, CMIP3) climate simulations (Supplementary Table 1) for the period 1979–99, for which both observations and simulations are available. All observed metrics show poleward displacement of the Northern Hemisphere tropical boundary by 0.2°–0.75° per decade. Although the JET and PMIN displacements are not significant at the 95% confidence level, the others are, and the combined metric ALL shows significant poleward displacement of 0.33° ± 0.12° per decade.

As shown previously[11], the CMIP3 models underestimate Northern Hemisphere expansion. However, we note that this failure is more evident in models that lack time-varying black carbon or ozone, wherein four of the five metrics actually move in the wrong direction. In models that do include one or both forcings, poleward displacement is robust across most indicators. In those with both forcings, ALL shows 0.14 ± 0.06° per decade—about half what is observed and significantly more than in models not including black carbon and ozone forcing (non-BC/O_3).

Expansion rates vary by season (Fig. 2, bottom). In the Northern Hemisphere, observed expansion is strongest in June–August (JJA) and September–November (SON) at 0.53° per decade and 0.58° per decade, respectively. Non-BC/O_3 CMIP3 models significantly underestimate expansion in these seasons. Models that include ozone, however, and those also including black carbon, simulate more expansion in JJA and SON. The impact of both forcings (BC + O_3), that is, the difference relative to non-BC/O_3, is greatest in JJA (0.24° per decade) followed by SON (0.20° per decade) and March–May (MAM) (0.15° per decade), with the smallest difference in December–February (DJF) (0.07° per decade). Such a seasonal cycle is similar to that of the observed trend, although expansion overall is still too small even in the BC + O_3 models. The impact of these forcings becomes more statistically significant in the models when examining a longer time period, 1970–1999 (Supplementary Fig. 3).

The preceding results are based on a relatively short time period, are probably affected by differences between the models that used different forcings, and were based on relatively crude aerosol treatments. We therefore conduct a suite of longer (1970–2009) simulations with a single climate model, the Community Atmosphere Model version 3 (CAM3)[18] of the National Center for Atmospheric Research (NCAR), equipped with new aerosol physics. We isolate the impact of a given forcing agent by comparing model runs with and without that agent.

The annual-mean observed displacement over 1979–2009, 0.38 ± 0.11° per decade, is about 15% stronger than for the shorter period and has approximately the same seasonal cycle (Fig. 3). The CAM 1979–2009, CAM 1970–2009, and CMIP3 BC + O_3 1979–1999 runs all produce similar counterpart expansions of about 0.14° per decade, with similar seasonal cycles. Therefore none of the important results are sensitive to the observing period.

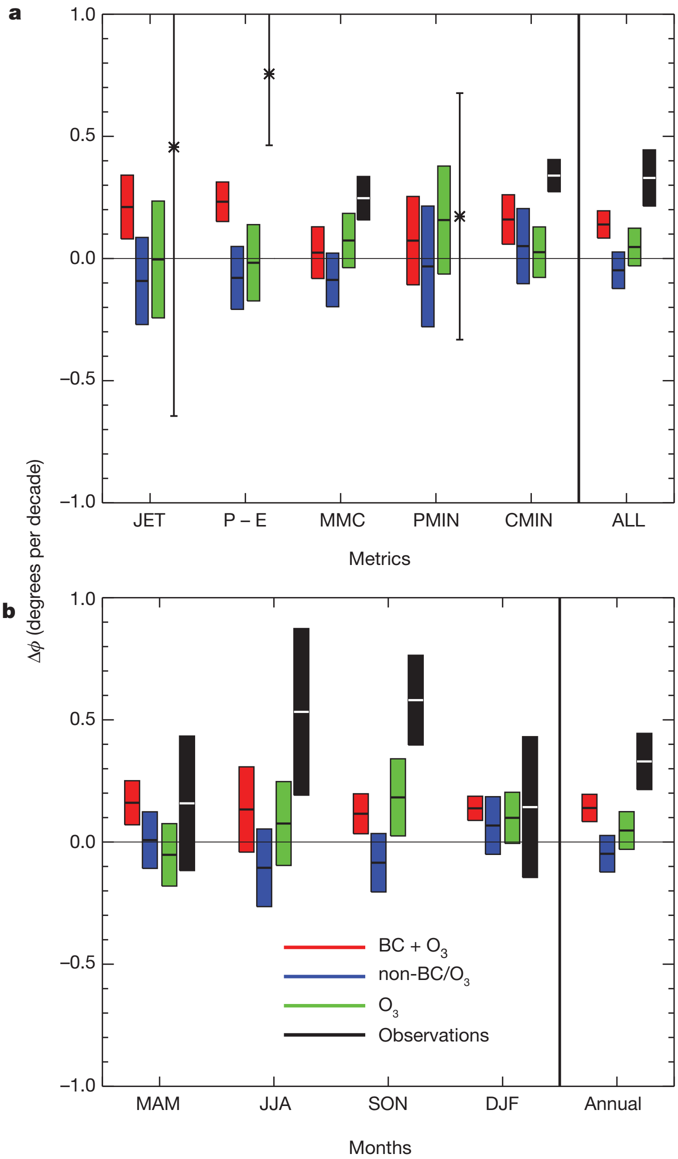


Figure 2 | Observed and modelled 1979–1999 Northern Hemisphere tropical expansion based on five metrics. **a**, Annual mean poleward displacement of each metric, as well as the combined ALL metric. **b**, Poleward displacement by season, based on ALL. CMIP3 models are grouped into nine that included time-varying black carbon and ozone (red); three that included time-varying ozone only (green); and six that included neither time-varying black carbon nor ozone (blue). Boxes show the mean response within each group (centre line) and its 2σ uncertainty. Observations are in black. In the case of one observational data set, trend uncertainty (whiskers) is estimated as the 95% confidence level according to a standard t-test.

These CAM simulations confirm the important role of time-varying black carbon and ozone in driving simulated expansion. In the non-BC/O_3 run, annual-mean expansion dropped to 0.04 ± 0.03° per decade, or about one-third of that with all forcings. Little if any of this is due to stratospheric ozone, given that almost the same result (0.06 ± 0.04° per decade) is obtained in the non-BC/tO_3 run, where tO_3 is tropospheric ozone. Although the latter result is not quite significantly different from the all-forcings response over 1979–2009, it is significantly different over the longer time period, and is robust across metrics (Supplementary Fig. 4). Moreover, as seen in the CMIP3 simulations, BC + tO_3 produces a much more realistic warm-season trend maximum; the JJA and MAM trend increases are statistically significant .

Black carbon and tropospheric ozone are negligible drivers of Southern Hemisphere expansion (Supplementary Fig. 5). Instead we

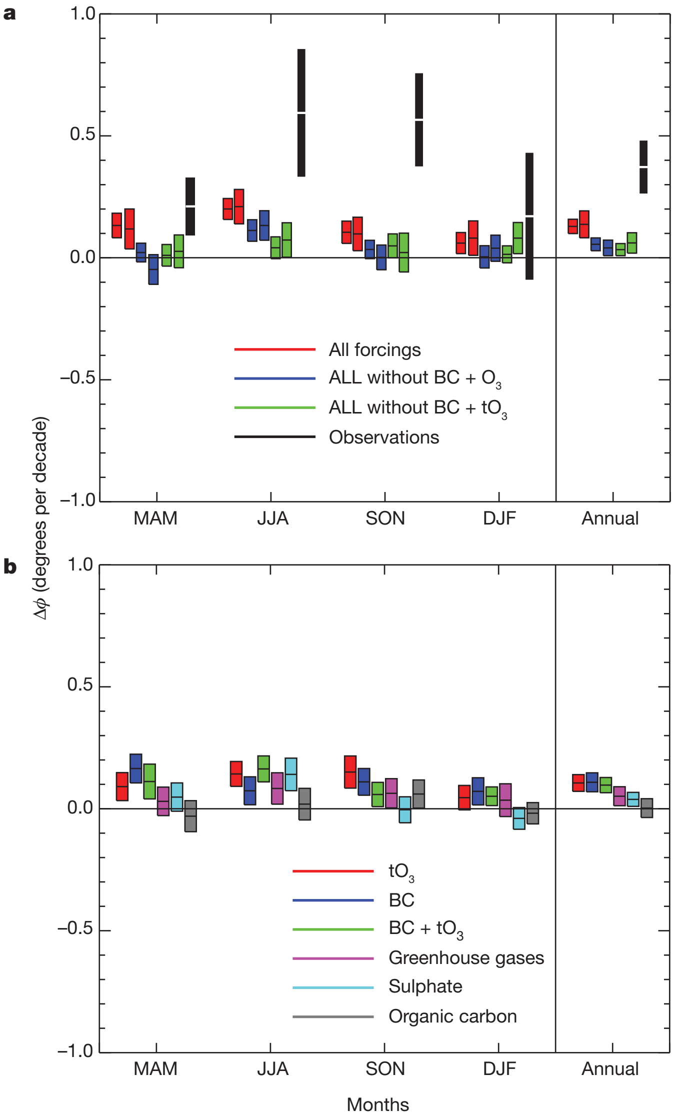


Figure 3 | Northern Hemisphere seasonal tropical expansion based on the combined ALL metric. **a**, CAM simulations for all forcings (red), all forcings except black carbon and ozone (blue); and all forcings except black carbon and tropospheric ozone (green). The first box in each like-coloured pair represents 1970–2009; the second box in each like-coloured pair represents 1979–2009. Observations (black) for 1979–2009 are also included. **b**, CAM individual forcing experiments for 1970–2009 showing the difference between the all forcings experiment and all forcings without tropospheric ozone (red), black carbon (blue), black carbon and tropospheric ozone (green), greenhouse gases (purple), sulphate (light blue) and organic carbon (grey). Boxes show the mean response (centre line) and its 2σ uncertainty.

find, as have previous studies[7–10], that stratospheric ozone depletion is the main driver there, particularly during the peak DJF expansion season.

We explored the role of various forcings more thoroughly by conducting additional ten-ensemble member simulations with CAM, individually omitting each of black carbon, tropospheric ozone, greenhouse gases, and the scattering aerosols—sulphate and organic carbon (Fig. 3b). In most seasons, greenhouse gases and heterogeneous warming agents (that is, black carbon and tropospheric ozone) push the Northern Hemisphere tropical boundary poleward. Over the year, greenhouse gases yield about 0.05° per decade, which is only significant for the longest time period, 1970–2009. Either tropospheric ozone, black carbon or their combination ($BC + tO_3$) each cause roughly twice this much expansion, ranging from 0.07° to 0.12° per decade, which is significant for both time periods (Supplementary Fig. 6 shows 1979–2009). The combined impact of black and organic carbon (which are generally emitted together) also falls in this range, showing that little of the expansion from black carbon is offset by organic carbon, which is non-absorbing. These results are again qualitatively robust across individual metrics, although less statistically significant (Supplementary Fig. 7). Each of the heterogeneous warming agents produces a more realistic seasonal trend cycle than do greenhouse gases. Decreases in scattering aerosols (due to declining mid-latitude sulphate) have not significantly contributed to Northern Hemisphere tropical expansion except during JJA, the time of maximum solar insolation.

We relate tropical expansion to a temperature index that compares mid-latitude tropospheric warming to that at other latitudes[12]. Warming of mid-latitudes relative to others displaces the maximum meridional climatological temperature gradient poleward. Geostrophic adjustment to this perturbed temperature gradient also implies a poleward shift of the tropospheric jet[12].

We consider a quantity called the 'expansion index': $2 \times \Delta T_{30-60} - (\Delta T_{0-30} + \Delta T_{60-90})$, where ΔT is the log-pressure (850–300 hPa) area-weighted temperature response in low (0°–30°), mid- (30°–60°), and high (60°–90°) latitudes[12]. As the expansion index becomes more positive, mid-latitude warming amplification dominates, and we expect more tropical expansion. Similarly, as the expansion index becomes less positive, mid-latitude cooling amplification dominates, and we expect less tropical expansion.

The above relationship helps to explain why black carbon and tropospheric ozone drive Northern Hemisphere tropical expansion (Supplementary Information and Supplementary Figs 8–14). Both agents heat primarily within the 30°–55° N latitude range. Although dynamical responses can cause the geographical patterns of applied heating and resulting warming to be quite different, in the zonal mean, mid-latitude heat input does appear to produce warming at roughly the heated latitudes[12], consistent with our results. Experiments with an alternative general circulation model, the Geophysical Fluid Dynamics Laboratory (GFDL) atmospheric model AM2.1 (ref. 19 and Supplementary Information), increase our confidence that this response to mid-latitude heating, and black carbon and tropospheric ozone in particular, is robust across different climate models, as well as different aerosol and ozone forcings.

This relationship also explains the seasonal cycle of the response, because both black carbon and tropospheric ozone warm primarily by absorbing solar radiation, which is far more abundant during summer. In our CAM simulations, atmospheric solar absorption by black carbon in the Northern Hemisphere mid-latitudes increases by more than a factor of three from DJF to JJA, 0.76 W m^{-2} versus 2.63 W m^{-2}. This results in about 0.05 K per decade more tropospheric warming in the Northern Hemisphere mid-latitudes during JJA compared to DJF. A similar variation results from tropospheric ozone (Supplementary Fig. 15).

Figure 4 quantifies the relationship between annual mean Northern Hemisphere tropical expansion and expansion index for 1970–2009. Climate forcing agents that yield a positive expansion index also yield a positive (poleward) displacement for most metrics, and vice versa. The corresponding correlation coefficient, over all experiments and metrics is 0.66, significant at the 99% confidence level. Correlations for the individual metrics are 0.86 for JET; 0.62 for $P - E$; 0.89 for MMC; 0.68 for PMIN; and 0.16 for CMIN.

Our analysis strongly suggests that recent Northern Hemisphere tropical expansion is driven mainly by black carbon and tropospheric ozone, with greenhouse gases playing a smaller part. Compared to observations, the magnitude of the simulated change is underestimated. This could be related to the aforementioned caveats with the observations, model deficiencies, or deficient black carbon aerosol forcing. The average top-of-the-atmosphere black carbon radiative forcing is 0.35 W m^{-2} for CMIP3 models[20], 0.43 W m^{-2} in our CAM simulations, and was reported as 0.25 W m^{-2} in a third suite of relatively

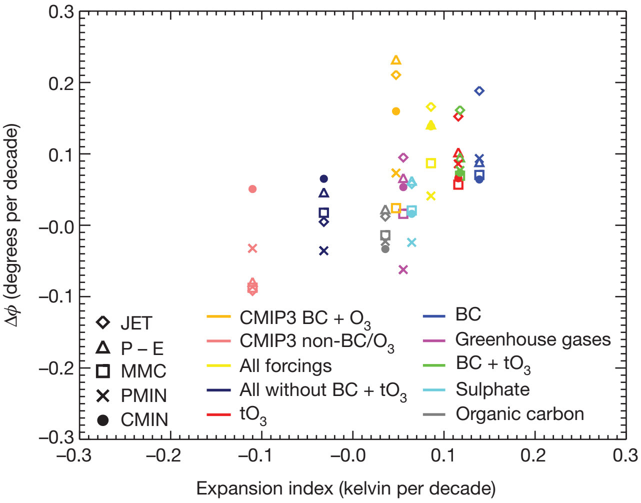


Figure 4 | Northern Hemisphere 1970–2009 annual mean tropical expansion for each metric versus the expansion index for CAM simulations. Also included are CMIP3 results from 1979–1999, stratified by BC + O_3 (orange) versus non-BC/O_3 (pink) models.

sophisticated models from the Aerosol Comparisons between Observations and Models (AeroCom)[21] project. However, recent observationally constrained estimates[22] range from 0.4 W m^{-2} to 1.2 W m^{-2}. Although others have inferred a similarly large magnitude of black carbon forcing[23,24], we emphasize that uncertainties still exist. It is also possible that the increase in black carbon emissions—particularly in southeast Asia—is underestimated, as has been inferred with CMIP3 aerosol emission inventories[25].

The upper end of the observed range of black carbon would reconcile the 2.5 factor shortfall in all-forcings expansion relative to observations, if responses varied linearly with forcing. However, our results show that responses are not always linear, and it still seems likely that models are insufficiently sensitive to these forcings. As long as this insensitivity applies equally to different forcings, our results point to anthropogenic pollutants other than CO_2 rather than global warming as the culprit in recent Northern Hemisphere tropical expansion. Emission controls on black carbon and ozone precursors would thus not only help mitigate global warming and improve human health[26], but could lessen the regional impacts of changes in large-scale Northern Hemisphere atmospheric circulation.

METHODS SUMMARY

CAM was run at T42 resolution coupled to the Community Land Model (CLM) version 3, a slab ocean-thermodynamic sea ice model, and the Snow, ICe and Aerosol Radiative (SNICAR) model[27,28] for the period 1970–2009. In addition to the usual natural forcings, our runs included radiatively active black carbon in the atmosphere and snow and an enhancement factor of 1.5 for solar absorption by coated hydrophilic particles[27,29]. Aerosol indirect effects were not included. Time-varying forcing followed the newly developed CMIP5 data set, including estimated concentrations of greenhouse gases and primary emissions of sulphur dioxide, black and organic carbon. Post-2005 emissions were derived from the average of the four representative concentration pathways and data are linearly interpolated to annual resolution. We show results from a ten-member ensemble with each member integrated from an independent initial condition based on a 30-year control simulation with constant 1970 forcing. Analysing the correlation between individual ensemble members' time series of several tropical expansion metrics supported the independence of each realization, because correlations on both annual and longer timescales ranged from −0.01 to 0.13.

Changes in tropical width were estimated by taking a least-squares trend of the seasonal or annual mean time series of each metric. The median of pairwise slopes regression yielded similar results. When multiple realizations (or observational data sets for a given metric) were available, trend uncertainty was estimated from the multiple realizations, as twice the standard error, $2 \times \frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the trends and n is the number of trends. In the case of one observational data set, trend uncertainty was estimated as the 95% confidence level according to a standard t-test[30].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 16 December 2011; accepted 29 March 2012.

1. Seidel, D. J., Fu, Q., Randel, W. J. & Reichler, T. J. Widening of the tropical belt in a changing climate. *Nature Geosci.* **1,** 21–24 (2008).
2. Hu, Y. & Fu, Q. Observed poleward expansion of the Hadley circulation since 1979. *Atmos. Chem. Phys.* **7,** 5229–5236 (2007).
3. Fu, Q., Johanson, C. M., Wallace, J. M. & Reichler, T. Enhanced mid-latitude tropospheric warming in satellite measurements. *Science* **312,** 1179 (2006).
4. Fu, Q. & Lin, P. Poleward shift of subtropical jets inferred from satellite-observed lower stratospheric temperatures. *J. Clim.* **24,** 5597–5603 (2011).
5. Zhou, Y. P., Xu, K.-M., Sud, Y. C. & Betts, A. K. Recent trends of the tropical hydrological cycle inferred from Global Precipitation Climatology Project and International Satellite Cloud Climatology Project data. *J. Geophys. Res.* **116,** D09101 (2011).
6. Bender, F., Ramanathan, V. & Tselioudis, G. Changes in extratropical storm track cloudiness 1983–2008: observational support for a poleward shift. *Clim. Dyn.* http://dx.doi.org/10.1007/s00382-011-1065-6 (2011).
7. Son, S.-W., Tandon, L. M., Polvani, L. M. & Waugh, D. W. Ozone hole and Southern Hemisphere climate change. *Geophys. Res. Lett.* **36,** L15705 (2009).
8. Polvani, L. M., Waugh, D. W., Correa, G. J. P. & Son, S.-W. Stratospheric ozone depletion: the main driver of twentieth-century atmospheric circulation changes in the Southern Hemisphere. *J. Clim.* **24,** 795–812 (2011).
9. Son, S.-W. *et al.* Impact of stratospheric ozone on Southern Hemisphere circulation change: a multimodel assessment. *J. Geophys. Res.* **115,** D00M07 (2010).
10. Kang, S. M., Polvani, L. M., Fyfe, J. C. & Sigmond, M. Impact of polar ozone depletion on subtropical precipitation. *Science* **332,** 951–954 (2011).
11. Johanson, C. M. & Fu, Q. Hadley cell widening: model simulations versus observations. *J. Clim.* **22,** 2713–2725 (2009).
12. Allen, R. J., Sherwood, S. C., Norris, J. R. & Zender, C. S. The equilibrium response to idealized thermal forcings in a comprehensive GCM: implications for recent tropical expansion. *Atmos. Chem. Phys. Discuss.* **11,** 31643–31688 (2011).
13. Lu, J., Vecchi, G. A. & Reichler, T. Expansion of the Hadley cell under global warming. *Geophys. Res. Lett.* **34,** L06805 (2007).
14. Lu, J., Deser, C. & Reichler, T. Cause of the widening of the tropical belt since 1958. *Geophys. Res. Lett.* **36,** L03803 (2009).
15. Allen, R. J. & Sherwood, S. C. The impact of natural versus anthropogenic aerosols on atmospheric circulation in the Community Atmosphere Model. *Clim. Dyn.* **36,** 1959–1978 (2011).
16. Ming, Y., Ramaswamy, V. & Chen, G. A model investigation of aerosol-induced changes in boreal winter extratropical circulation. *J. Clim.* **24,** 6077–6091 (2011).
17. Quaas, J., Boucher, O., Bellouin, N. & Kinne, S. Satellite-based estimate of the direct and indirect aerosol climate forcing. *J. Geophys. Res.* **113,** D05204 (2008).
18. Collins, W. D. *et al. Description of the NCAR Community Atmosphere Model (CAM3).* Technical Report NCAR/TN-464+STR (National Center for Atmospheric Research, 2004).
19. The GFDL Global Atmospheric Model Development Team.. The new GFDL global atmosphere and land model AM2–LM2: evaluation with prescribed SST simulations. *J. Clim.* **17,** 4641–4673 (2004).
20. Forster, P. *et al.* In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon, S. *et al.*) Ch. 2 130–234 (Cambridge University Press, 2007).
21. Schulz, M. *et al.* Radiative forcing by aerosols as derived from the AeroCom present-day and pre-industrial simulations. *Atmos. Chem. Phys.* **6,** 5225–5246 (2006).
22. Ramanathan, V. & Carmichael, G. Global and regional climate changes due to black carbon. *Nature Geosci.* **1,** 221–227 (2008).
23. Sato, M. *et al.* Global atmospheric black carbon inferred from AERONET. *Proc. Natl Acad. Sci. USA* **100,** 6319–6324 (2003).
24. Chung, S. H. & Seinfeld, J. H. Global distribution and climate forcing of carbonaceous aerosols. *J. Geophys. Res.* **107,** D19 (2002).
25. Dwyer, J. G., Norris, J. R. & Ruckstuhl, C. Do climate models reproduce the observed solar dimming and brightening over China and Japan? *J. Geophys. Res.* **115,** D00K08 (2010).
26. Shindell, D. *et al.* Simultaneously mitigating near-term climate change and improving human health and food security. *Science* **335,** 183–189 (2012).
27. Flanner, M. G., Zender, C. S., Randerson, J. T. & Rasch, P. J. Present-day climate forcing and response from black carbon in snow. *J. Geophys. Res.* **112,** D11202 (2007).
28. Flanner, M. G. *et al.* Springtime warming and reduced snow cover from carbonaceous particles. *Atmos. Chem. Phys.* **9,** 2481–2497 (2009).
29. Bond, T. C. & Bergstrom, R. W. Light absorption by carbonaceous aerosols: an investigative review. *Aerosol Sci. Technol.* **40,** 27–67 (2006).
30. Wilks, D. S. *Statistical Methods in the Atmospheric Sciences* (Academic Press, 2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This study was funded by R.J.A.'s University of California at Riverside initial complement. We acknowledge the individual modelling groups, the

Program for Climate Model Diagnosis and Intercomparison and the Working Group on Coupled Modeling of the World Climate Research Programme (WCRP) for their part in making available the WCRP CMIP3 multimodel data set. Support of this data set is provided by the Office of Science, US Department of Energy.

Author Contributions R.J.A. conceived the project, designed the study, carried out all data analysis and wrote the manuscript. S.C.S. advised on methods and interpretation, and assisted in the writing of the manuscript. J.N. provided homogenized cloud data and assisted in experimental design. C.S.Z. assisted with CAM experiments, including the SNICAR model.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.J.A. (rjallen@ucr.edu).

METHODS

Our black carbon radiative forcing is computed interactively at each time step as the difference in fluxes with all species present and all species except black carbon. Thus, our 0.43 W m^{-2} top-of-the-atmosphere radiative forcing is a measure of the instantaneous forcing, which for aerosols is a close approximation of the (adjusted) radiative forcing used by CMIP3 models. CAM alterations, including the SNICAR model and modification of black carbon optical properties to account for enhanced solar absorption by coated hydrophilic particles, are described elsewhere[27,28].

Observational data comes from several sources, including the Global Precipitation Climatology Project version 2.2 (GPCP)[31] the Integrated Global Radiosonde Archive (IGRA)[32] and five reanalyses[33–37] for MMC calculations. Cloud cover observations, which span July 1983 to June 2008 only, come from two recently homogenized satellite data sets[38] based on the International Satellite Cloud Climatology Project (ISCCP)[39,40] and the Advanced Very High Resolution Radiometer (AVHRR) Pathfinder Atmospheres Extended (PATMOS-x)[41,42]. Our P − E estimate is based on precipitation from GPCP and evaporation from the Woods Hole Oceanographic Institution (WHOI) Objectively Analyzed air-sea Flux (OAFlux) project[43]. Because the WHOI OAFlux data are over ocean only, our P − E estimate is for the global oceans.

For the IGRA jet analysis, monthly mean zonal wind data are used at stations with 75% valid years, where a valid year has four valid seasons and a valid season has two of three valid months. This is required at all tropospheric pressure levels (850 hPa, 700 hPa, 500 hPa, 400 hPa, 300 hPa). To minimize trend errors, we also required a station to possess eight valid years in the first and last decade. Data from both 00Z and 12Z are used. This resulted in 273 12Z and 300 00Z IGRA stations for 1979–1999 and 247 12Z and 281 00Z stations for 1979–2009, most of which are in the Northern Hemisphere. Data are gridded to $5^\circ \times 10^\circ$ resolution by assigning each station's monthly zonal wind to the nearest grid point without interpolation. When more than one station matched the same grid point, that grid point's value is estimated as the average of the available station values. Sub-sampling the CMIP3 models at the IGRA station locations yielded similar results, but sub-sampling decreases the ensemble mean jet displacement from $0.11 \pm 0.10^\circ$ per decade to $0.03 \pm 0.16^\circ$ per decade.

Our jet-based measure of tropical width is based on locating the 'sides' of the jet using the 75th percentile of monthly mean zonal wind for each tropospheric (850–300 hPa) pressure level[12]. The 75th percentile is estimated by sorting the monthly mean zonal wind—for each hemisphere and tropospheric pressure level—from low to high and taking the $0.75 \times (N+1)$ value, where N is the number of zonal wind values (that is, latitudes). Taking the midpoint and averaging over pressure levels yields a time series of monthly jet locations for each hemisphere.

Displacements for all metrics are estimated by first smoothing the zonal monthly mean of the appropriate field(s) and interpolating to 0.5° resolution using cubic splines. Smoothing was performed by taking a running mean over about 10° of latitude. However, nearly identical results are obtained without interpolating.

For trend uncertainty calculations (in the case of one observational metric), the influence of serial correlation is accounted for by using the effective sample size, $n(1-r_1)(1+r_1)^{-1}$, where n is the number of years and r_1 is the lag-1 autocorrelation coefficient[30].

When multiple realizations (or observational data sets for a given metric) are available, trend uncertainty is estimated as twice the standard error. These 2σ ranges are approximate, given that we have a ten-member ensemble and cannot confirm that the trends are Gaussian-distributed. However, ten ensemble members is relatively large for such a study and is the largest we could manage given the high computational costs of running nine forcing scenarios.

31. Adler, R. *et al.* The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeorol.* **4,** 1147–1167 (2003).
32. Durre, I., Vose, R. S. & Wuertz, D. M. Overview of the Integrated Global Radiosonde Archive. *J. Clim.* **19,** 53–68 (2006).
33. Kalnay, E. *et al.* The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77,** 437–471 (1996).
34. Kanamitsu, M. *et al.* NCEP-DOE AMIP-II Reanalysis (R-2). *Bull. Am. Meteorol. Soc.* **83,** 1631–1643 (2002).
35. Uppala, S. *et al.* The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131,** 2961–3012 (2005).
36. Rienecker, M. M. *et al.* MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Clim.* **24,** 3624–3648 (2011).
37. Saha, S. *et al.* The NCEP Climate Forecast System Reanalysis. *Bull. Am. Meteorol. Soc.* **91,** 1015–1057 (2010).
38. Clement, A., Burgman, R. & Norris, J. R. Observational and model evidence for positive low-level cloud feedback. *Science* **325,** 460–464 (2009).
39. Rossow, W. B. & Schiffer, R. A. ISCCP cloud data products. *Bull. Am. Meteorol. Soc.* **72,** 2–20 (1991).
40. Rossow, W. B. & Schiffer, R. A. Advances in understanding clouds from ISCCP. *Bull. Am. Meteorol. Soc.* **80,** 2261–2287 (1999).
41. Jacobowitz, H. *et al.* The Advanced Very High Resolution Radiometer Pathfinder Atmosphere (PATMOS) climate data set: a resource for climate research. *Bull. Am. Meteorol. Soc.* **84,** 785–793 (2003).
42. Pavolonis, M., Heidinger, A. & Uttal, T. Daytime global cloud typing from AVHRR and VIRS: algorithm description, validation, and comparisons. *J. Appl. Meteorol.* **44,** 804–826 (2005).
43. Yu, L. & Weller, R. A. Objectively analyzed air-sea heat fluxes for the global ice-free oceans (1981–2005). *Bull. Am. Meteorol. Soc.* **88,** 527–539 (2007).

LETTER

doi:10.1038/nature11031

Thermal and electrical conductivity of iron at Earth's core conditions

Monica Pozzo[1], Chris Davies[2], David Gubbins[2,3] & Dario Alfè[1,4]

The Earth acts as a gigantic heat engine driven by the decay of radiogenic isotopes and slow cooling, which gives rise to plate tectonics, volcanoes and mountain building. Another key product is the geomagnetic field, generated in the liquid iron core by a dynamo running on heat released by cooling and freezing (as the solid inner core grows), and on chemical convection (due to light elements expelled from the liquid on freezing). The power supplied to the geodynamo, measured by the heat flux across the core–mantle boundary (CMB), places constraints on Earth's evolution[1]. Estimates of CMB heat flux[2–5] depend on properties of iron mixtures under the extreme pressure and temperature conditions in the core, most critically on the thermal and electrical conductivities. These quantities remain poorly known because of inherent experimental and theoretical difficulties. Here we use density functional theory to compute these conductivities in liquid iron mixtures at core conditions from first principles—unlike previous estimates, which relied on extrapolations. The mixtures of iron, oxygen, sulphur and silicon are taken from earlier work[6] and fit the seismologically determined core density and inner-core boundary density jump[7,8]. We find both conductivities to be two to three times higher than estimates in current use. The changes are so large that core thermal histories and power requirements need to be reassessed. New estimates indicate that the adiabatic heat flux is 15 to 16 terawatts at the CMB, higher than present estimates of CMB heat flux based on mantle convection[1]; the top of the core must be thermally stratified and any convection in the upper core must be driven by chemical convection against the adverse thermal buoyancy or lateral variations in CMB heat flow. Power for the geodynamo is greatly restricted, and future models of mantle evolution will need to incorporate a high CMB heat flux and explain the recent formation of the inner core.

First principles calculations of transport properties based on density functional theory (DFT) have been used in the past for a number of materials (see, for example, refs 9, 10). Recently, increased computer power has facilitated simulations of large systems, allowing the problem of the size of the simulation cell to be addressed: this can be a serious problem for the electrical conductivity, σ (ref. 11). Here we report a series of calculations of the electrical and thermal conductivity (k) of iron at Earth's core conditions, using DFT. We previously used these methods to compute an extensive number of thermodynamic properties of iron and iron alloys, including the whole melting curve of iron in the pressure range 50–400 GPa (refs 12, 13) and the chemical potentials of oxygen, sulphur and silicon in solid and liquid iron at inner core boundary (ICB) conditions, which we used to place constraints on core composition[6]. Recently, we computed the conductivity of iron at ambient conditions, and obtained values in very good agreement with experiments[14].

We calculated three adiabatic temperature–pressure profiles (adiabats) for the core; to do this, we assumed three different possible temperatures at the ICB, and followed the line of constant entropy as the pressure was reduced to that of the CMB. The ICB temperatures were: 6,350 K (the melting temperature of pure iron)[13], 5,700 K (the melting temperature of a mixture of iron with 10% Si and 8% O, corresponding to an inner-core density jump $\Delta\rho = 0.6\ \mathrm{g\,cm^{-3}}$)[6] and 5,500 K (the melting temperature of a mixture of iron with 8% Si and 13% O, corresponding to $\Delta\rho = 0.8\ \mathrm{g\,cm^{-3}}$)[6]. Then we calculated the electrical and thermal conductivity of iron at seven positions on these three adiabats. Our results are reported in Fig. 1, and show a smooth variation of these parameters in the core; σ only varies by $\sim$13% between the ICB and the CMB, and it is almost the same for all adiabats. A recent shock wave experiment[15] reported $\sigma = 0.765 \times 10^6\ \Omega^{-1}\,\mathrm{m}^{-1}$ for pure iron at 208 GPa, and an older shock wave measurement[16] reported $\sigma = 1.48 \times 10^6\ \Omega^{-1}\,\mathrm{m}^{-1}$ at 140 GPa. Our values are closer to the latter. There is a larger variation in k, as implied by the Wiedemann–Franz law (which relates the thermal and electrical conductivity through $L = k/\sigma T$), which we found to be closely followed throughout the core with a Lorenz parameter $L = (2.48–2.5) \times 10^{-8}\ \mathrm{W\,\Omega\,K^{-2}}$. The ionic contribution to k was calculated using the classical potential used as a reference system in ref. 12, which was shown to describe very accurately the energetics of the system and the structural and dynamical properties of liquid iron at Earth's core conditions. We found that the ionic contribution is only between 2.5 and $4\ \mathrm{W\,m^{-1}\,K^{-1}}$ on the adiabat, which is negligible compared to the electronic contribution, as expected.

The estimates of k (Fig. 1) are substantially larger than previously used in the geophysical literature, approximately doubling the heat conducted down the adiabatic gradient in the core and halving the power to drive a dynamo generating the same magnetic field. These considerations demand a revision of the power requirements for the geodynamo. The conductivities for liquid mixtures appropriate to the outer core are likely to be smaller than for pure iron, preliminary calculations suggesting about 30% lower, a smaller difference than that found in previous work[17], but in close agreement with extrapolations obtained from recent diamond-anvil-cell experiments, which reported a value in the range 90–130 $\mathrm{W\,m^{-1}\,K^{-1}}$ at the top of the outer core[18]. Our values are also in broad agreement with recently reported DFT calculations[19].

We focus on estimates for the two mixtures above, corresponding to ICB density jumps $0.6\ \mathrm{g\,cm^{-3}}$ (ref. 8) and $0.8\ \mathrm{g\,cm^{-3}}$ (ref. 7). There is relatively little effect on the conductivities in the two cases, because any additional O in the outer core must be balanced by less S or Si to maintain the mass of the whole core, which is well constrained. The larger density jump gives a higher O content, more gravitational energy, a lower ICB temperature and lower adiabatic gradient: it therefore favours compositional over thermal convection. The relevant values are given in Table 1.

We estimate power requirements for the dynamo using the model described in a previous study (ref. 5, and Methods). Neglecting small effects, the total CMB heat flux, Q_{CMB}, is the sum of terms proportional to either the CMB cooling rate, $\mathrm{d}T_0/\mathrm{d}t$, or the amount of radiogenic heating, h: $Q_{\mathrm{CMB}} = Q_s + Q_L + Q_g + Q_r$, where the terms

[1]Department of Earth Sciences, and Thomas Young Centre at UCL, UCL, Gower Street, London WC1E 6BT, UK. [2]School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK. [3]Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California at San Diego, 9500 Gilman Drive no. 0225, La Jolla, California 92093-0225, USA. [4]Department of Physics and Astronomy, and London Centre for Nanotechnology, UCL, Gower Street, London WC1E 6BT, UK.

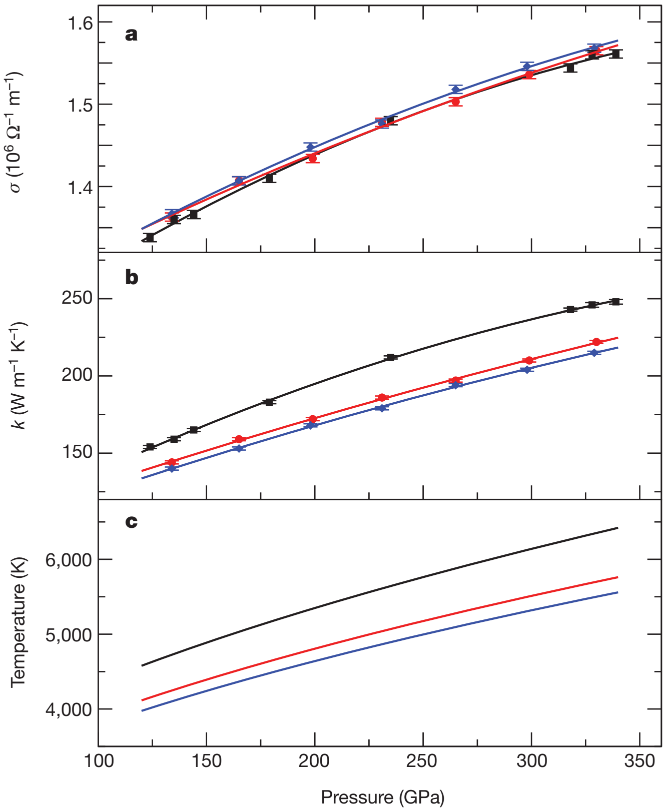


Figure 1 | Electrical and thermal conductivity of iron at Earth's outer core conditions. a–c, Electrical conductivity, σ (**a**), and electronic component of thermal conductivity, k (**b**), of pure iron corresponding to the three outer-core adiabatic profiles (adiabats) displayed in **c**. Black lines, adiabat corresponding to the melting temperature of pure iron at ICB pressure; red lines, that of the mixture containing 10% Si and 8% O; and blue lines, that of the mixture with 8% Si and 13% O. Lines are quadratic fits to the first principles raw data (symbols). Error bars (2 s.d.) are estimated from the scattering of the data obtained from 40 statistical independent configurations. Results are obtained with cells including 157 atoms and the single **k**-point (1/4,1/4,1/4), which are sufficient to obtain convergence within less than 1%.

on the right-hand side represent respectively the effects of secular cooling, latent heat, gravitational energy and radiogenic heating. The cooling rate, expressed in degrees per billion years, can be varied together with the radiogenic heating to produce some desired outcome: a fixed mantle heat flux, a marginal dynamo (no entropy left for ohmic dissipation, E_σ), or a primordial inner core (by decreasing the cooling rate and increasing the radiogenic heating). Results for a suite of 11 models are shown in Table 2.

Model 1 fails as a dynamo. There is an entropy deficit, meaning the assumption that the whole core can convect is incorrect—the temperature gradient must fall below the adiabat to balance the entropy equation. A dynamo might still be possible with a large part of the core completely stratified. Model 2 demonstrates the efficiency of compositional convection: the entropy is greatly increased compared to model 1 with no change in cooling rate and little increase in heat flux; the dynamo is now marginal. Model 3 has an increased cooling rate and consequent younger inner core to demonstrate what is required for a marginal dynamo with $\Delta\rho = 0.6\,\mathrm{g\,cm^{-3}}$. Models 4 and 5 have cooling rates that make the CMB thermally neutral; the CMB heat flux is equal to that conducted down the adiabat. Models 6 and 7 have some radiogenic heating and the original cooling rate and operate as dynamos, although they are still thermally stable at the top of the core. Models 8–11 have cooling rates that yield old inner-core ages, 3.5 and 4.5 Gyr, and the radiogenic heating has been adjusted to make a marginal dynamo. They are also thermally stable at the top of the core.

We estimate stable layer thicknesses by computing the radial variation of thermal and compositional gradients for each model using the equations of a previous study (ref. 20, Methods), which are derived from the equations of core energetics[5]. To compare thermal and chemical gradients, we multiply the latter by the ratio of compositional and thermal expansion coefficients α_c/α_T, thereby converting compositional effects into equivalent thermal effects. The base of the stable layer is defined as the point where the stabilizing adiabatic gradient, T'_a, crosses the combined destabilizing gradient, $T' = T'_L + T'_s + T'_c + T'_r$, where the terms represent respectively latent heat, secular cooling, compositional buoyancy and radiogenic heating.

Stable layer thicknesses are hundreds of kilometres in all models except those with cooling rates that are so rapid as to make the inner core too young; without compositional buoyancy the layers in all models except 4 and 5 span half the core (Table 2). Radiogenic heating thins the layers for the same cooling rate. Profiles of stabilizing and destabilizing gradients (Fig. 2) show that destabilizing gradients are greatest at depth, but much reduced compared to previous models[20] because they each depend on a factor $1/k$. The thermal conductivity increases by 50% across the core, increasing the heat conducted down the adiabat at depth and further reducing the power available to drive convection near the base of the outer core. Combined thermochemical profiles suggest that compositional buoyancy near the top of core is not strong enough to drive convection against the adverse temperature gradient.

Stable layers could be thinned or partially disturbed by convection, through penetration or instability, or some other effect not included in our simple model. A potentially more effective mechanism for inducing vertical mixing near the CMB is through lateral variations in CMB heat flux, which can drive motions without having to overcome the gravitational force. The presence of lateral variations makes the relevant heat flux for core mixing the maximum at the CMB[21], which could be as much as 10 times the average[22]; this does not influence dynamo entropy calculations but does allow magnetic flux to be carried to the surface in regions of cold mantle, as is observed[23].

As well as raising k, our calculations also raise σ to about twice the current estimate. Two important quantities depend on σ: the magnetic diffusion time (the time taken for the slowest decaying dipole mode to fall by a factor of e in the absence of a dynamo) and the magnetic Reynolds number Rm, which measures the rate of generation of

Table 1 | Parameters used to estimate power requirements for the geodynamo

$\Delta\rho$	T_{ICB}	T_{CMB}	k_{ICB}	k_{CMB}	σ_{ICB} ($\times10^6$)	σ_{CMB} ($\times10^6$)	O	S/Si
0.6	5,700	4,186	150 (223)	100 (144)	1.25 (1.56)	1.11(1.36)	8	10
0.8	5,500	4,039	150 (215)	100 (140)	1.24(1.57)	1.11(1.37)	13	8

Values in parenthesis are for pure iron, other values are approximations for core mixtures. Units are $\mathrm{g\,cm^{-3}}$ for the ICB density jump, $\Delta\rho$; K for the temperatures, T; $\mathrm{W\,m^{-1}\,K^{-1}}$ for the thermal conductivity, k; $\Omega^{-1}\,\mathrm{m^{-1}}$ for the electrical conductivity, σ; % for molar concentrations.

Table 2 | Heat flux and entropy for various models of cooling and radiogenic heating

Model	$\Delta\rho$	dT_0/dt	h	Q_{ad}	Q_{CMB}	IC age	E_σ	Δ
1	0.6	46	0	15.7	5.8	0.9	−111	1,022
2	0.8	46	0	15.2	6.1	1.0	5	826
3	0.6	57	0	15.7	7.2	0.7	−2	833
4	0.6	123	0	15.7	15.6	0.3	652	110
5	0.8	115	0	15.2	15.2	0.4	865	0
6	0.6	46	3.0	15.7	11.7	0.9	85	659
7	0.8	46	3.0	15.2	11.9	1.0	208	468
8	0.6	11.2	6.8	15.7	14.7	3.5	−3	1,257
9	0.6	8.7	6.9	15.7	14.5	4.5	−1	1,472
10	0.8	12.2	6.3	15.2	13.7	3.5	4	1,000
11	0.8	9.5	6.6	15.2	14.1	4.5	2	1,128

Here $\Delta\rho$ is the density jump at the ICB in $\mathrm{g\,cm^{-3}}$; dT_0/dt the cooling rate of the CMB in $\mathrm{K\,Gyr^{-1}}$; h the radiogenic heat source in $\mathrm{pW\,kg^{-1}}$; $Q_{ad} = -4\pi k(dT_{ad}/dr)$ is the heat conducted down the adiabat in TW where dT_{ad}/dr is the adiabatic gradient; Q_{CMB} is the heat flux across the CMB in TW; E_σ is the entropy available for the dynamo and other diffusive processes in $\mathrm{MW\,K^{-1}}$. Inner core (IC) age is shown in Gyr; stable layer thicknesses, Δ, are given in kilometres below the CMB.

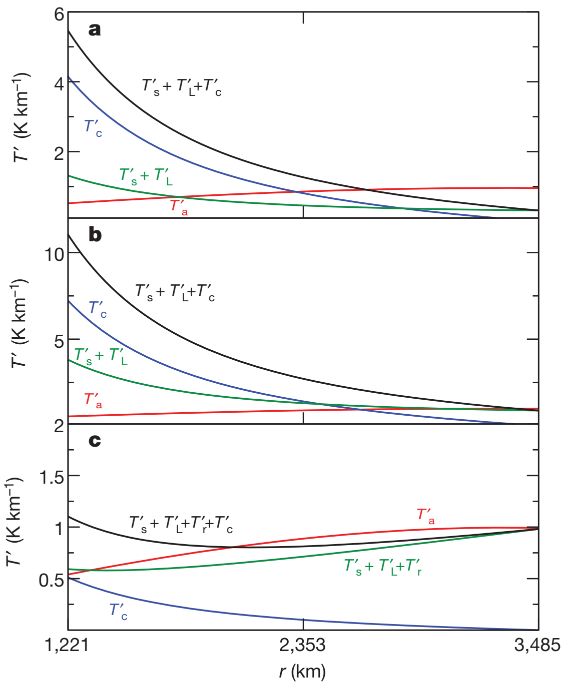


Figure 2 | Stabilizing and destabilizing gradients for three core energetics models. Equivalent temperature gradients, T', plotted against radius for three core evolution models. The stabilizing gradient is due to conduction down the adiabat, T'_a (red lines). Compositional buoyancy is denoted by T'_c (blue lines), latent heat by T'_L, secular cooling by T'_s and radiogenic heating by T'_r. The total destabilizing thermal gradient is represented by the green lines; total destabilizing thermochemical gradients are represented by black lines. Three models from Table 2 are shown: **a**, model 2 ($\Delta\rho = 0.8\,\mathrm{g\,cm^{-3}}$, $\mathrm{d}T_o/\mathrm{d}t = 46\,\mathrm{K\,Gyr^{-1}}$ and $h = 0$); **b**, model 4 ($\Delta\rho = 0.6\,\mathrm{g\,cm^{-3}}$, $\mathrm{d}T_o/\mathrm{d}t = 123\,\mathrm{K\,Gyr^{-1}}$ and $h = 0$); **c**, model 9 ($\Delta\rho = 0.6\,\mathrm{g\,cm^{-3}}$, $\mathrm{d}T_o/\mathrm{d}t = 8.7\,\mathrm{K\,Gyr^{-1}}$ and $h = 6.9\,\mathrm{pW\,kg^{-1}}$).

magnetic energy by a given flow. The magnetic diffusion time is increased to about 50 kyr. This may have significant implications for the theory of the secular variation: it makes the frozen flux approximation more accurate and lengthens the timescale of all diffusion-dominated processes, including polarity reversals. If current estimates of Rm are appropriate for the core[24], the increased conductivity implies that the geodynamo can operate on slower fluid flows and less input power from thermal and compositional convection.

Revised estimates of σ and k calculated directly at core conditions have fundamental consequences for the thermochemical evolution of the deep Earth. New estimates of the power requirements for the geodynamo suggest a CMB heat flux in the upper range of what is considered reasonable for mantle convection unless very marginal dynamo action can be sustained, while a primordial inner core is only possible with a significant concentration of radiogenic elements in the core. There are objections to a high CMB heat flux and also to radiogenic heating in the core[25–27], but one of the two seems inevitable if we are to have a dynamo. If the inner core is young, these high values of conductivity provide further problems with maintaining a purely thermally driven dynamo. A thermally stratified layer at the top of the core also appears inevitable. Viable thermal history models that produce thin stable layers and an inner core of age ~1 Gyr are likely to require a fairly rapid cooling rate and some radiogenic heating. The presence of a stable layer, and the effects associated with an increased electrical conductivity, have significant implications for our understanding of the geomagnetic secular variation.

METHODS SUMMARY

Calculations were performed using DFT with the same technical parameters used in refs 6, 12–14. We used the VASP code[28], PAW potentials[29,30] with $4s^1 3d^7$ valence configuration, the Perdew–Wang[31] functional, a plane wave cut-off of 293 eV, and single particle orbitals were occupied according to Fermi–Dirac statistics. We tested the effect on the conductivity of the inclusion in valence of semi-core $3s$ and $3p$ states; we found that, as in the zero pressure case[14], this effect is completely negligible.

The electrical conductivity and the electrical component of the thermal conductivity have been calculated using the Kubo–Greenwood formula and the Chester–Thellung–Kubo–Greenwood formula as implemented in VASP[32]. Because of the low mass of the electrons compared to the ions, the conductivities may be calculated by assuming frozen ionic configurations, and averaging over a sufficiently large set representing the typical distribution of the ions at the pressures and temperatures of interest.

Molecular dynamics simulations were performed in the canonical ensemble using cubic simulation cells with 157 atoms and the Γ point, a time step of 1 fs, and an efficient extrapolation of the charge density which speeds up the simulations by roughly a factor of two (ref. 33). Each state point was simulated for at least 6 ps, from which we discarded the first picosecond to allow for equilibration and used the last 5 ps to extract 40 configurations separated by 0.125 ps. This time interval is roughly two times longer than the correlation time, and therefore the configurations are statistically independent from each other. Because of the high temperatures involved, the conductivities converge quickly with respect to **k**-point sampling and size of the simulation cell[14], and we found that with a 157-atom cells and the single **k**-point (1/4,1/4/,1/4) the results are converged to better than 1%.

The ionic component of the thermal conductivity was calculated using the Green–Kubo formula.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 November 2011; accepted 7 March 2012.
Published online 11 April 2012.

1. Lay, T., Hernlund, J. & Buffett, B. Core-mantle boundary heat flow. *Nature Geosci.* **1,** 25–32 (2008).
2. Labrosse, S., Poirier, J.-P. & Le Mouël, J.-L. On cooling of the Earth's core. *Phys. Earth Planet. Inter.* **99,** 1–17 (1997).
3. Buffett, B., Garnero, E. & Jeanloz, R. Sediments at the top of Earth's core. *Science* **290,** 1338–1342 (2000).
4. Lister, J. R. & Buffett, B. A. The strength and efficiency of thermal and compositional convection in the geodynamo. *Phys. Earth Planet. Inter.* **91,** 17–30 (1995).
5. Gubbins, D., Alfè, D., Masters, T. G. & Price, D. Gross thermodynamics of 2-component core convection. *Geophys. J. Int.* **157,** 1407–1414 (2004).
6. Alfè, D., Gillan, M. J. & Price, G. D. Temperature and composition of the Earth's core. *Contemp. Phys.* **48,** 63–80 (2007).
7. Masters, T. G. & Gubbins, D. On the resolution of density within the Earth. *Phys. Earth Planet. Inter.* **140,** 159–167 (2003).
8. Dziewonski, A. M. & Anderson, D. L. Preliminary Reference Earth Model. *Phys. Earth Planet. Inter.* **25,** 297–356 (1981).
9. Silvestrelli, P. L., Alavi, A. & Parrinello, M. Electrical conductivity calculation in ab initio simulations of metals: application to liquid sodium. *Phys. Rev. B* **55,** 15515–15522 (1997).
10. Mattsson, T. R. & Desjarlais, M. P. Phase diagram and electrical conductivity of high energy density water from density functional theory. *Phys. Rev. Lett.* **97,** 017801 (2007).
11. Pozzo, M., Desjarlais, M. P. & Alfè, D. Electrical and thermal conductivity of liquid sodium from first principles calculations. *Phys. Rev. B* **84,** 054203 (2011).
12. Alfè, D., Gillan, M. J. & Price, G. D. The melting curve of iron at the pressures of the Earth's core conditions. *Nature* **401,** 462–464 (1999).
13. Alfè, D. Temperature of the inner-core boundary of the Earth: melting of iron at high pressure from first-principles coexistence simulations. *Phys. Rev. B* **79,** 060101(R) (2009).
14. Alfè, D., Pozzo, M. & Desjarlais, M. P. Lattice electrical resistivity of magnetic body-centred cubic iron from first principles calculations. *Phys. Rev. B* **85,** 024102 (2012).
15. Bi, Y., Tan, H. & Jing, F. Electrical conductivity of iron under shock compression up to 200 GPa. *J. Phys. Condens. Matter* **14,** 10849–10854 (2002).
16. Keeler, R. N. & Royce, E. B. in *Physics of High Energy Density* (eds Caldirola, P. & Knoepfel, H.) 106–125 (Proc. Int. Sch. Phys. Enrico Fermi Vol. 48, 1971).
17. Stacey, F. D. & Anderson, O. L. Electrical and thermal conductivities of Fe–Ni–Si alloy under core conditions. *Phys. Earth Planet. Inter.* **124,** 153–162 (2001).
18. Hirose, K. Gomi, H. Ohta, K., Labrosse, S. & Hernlund, J. The high conductivity of iron and thermal evolution of the Earth's core. *Mineral. Mag.* **75,** 1027 (2011).
19. de Koker, N., Steinle-Neumann, G. & Vlcek, V. Electrical resistivity and thermal conductivity of liquid Fe alloys at high P and T, and heat flux in Earth's core. *Proc. Natl Acad. Sci.* **109,** 4070–4073 (2012).
20. Davies, C. J. & Gubbins, D. A buoyancy profile for the Earth's core. *Geophys. J. Int.* **187,** 549–563 (2011).
21. Olson, P. in *Earth's Core and Lower Mantle* (eds Jones, C., Soward, A. & Zhang, K.) 1–49 (Taylor and Francis, London, 2000).

22. Nakagawa, T., &. Tackley, P. J. Lateral variations in CMB heat flux and deep mantle seismic velocity caused by a thermal-chemical-phase boundary layer in 3D spherical convection. *Earth Planet. Sci. Lett.* **271,** 348–358 (2008).
23. Jackson, A., Jonkers, A. R. T. & Walker, M. R. Four centuries of geomagnetic secular variation from historical records. *Phil. Trans. R. Soc. Lond. B* **358,** 957–990 (2000).
24. Gubbins, D. in *Encyclopedia of Geomagnetism and Paleomagnetism* (eds Gubbins, D. & Herrero-Bervera, E.) 287–300 (Springer, 2007).
25. Davies, G. Topography: a robust constraint on mantle fluxes. *Chem. Geol.* **145,** 479–489 (1998).
26. Davies, G. Mantle regulation of core cooling: a geodynamo without core radioactivity? *Phys. Earth Planet. Inter.* **160,** 215–229 (2007).
27. McDonough, W. in *Treatise on Geochemistry* Vol. 2 (ed. Carlson, R. W.) 547–568 (Elsevier, 2003).
28. Kresse, G. & Furthmuller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6,** 15–50 (1996).
29. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50,** 17953–17979 (1994).
30. Kresse, G&. Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59,** 1758–1775 (1999).
31. Wang, Y. & Perdew, J. P. Correlation hole of the spin-polarized electron gas, with exact small-wave-vector and high-density scaling. *Phys. Rev. B* **44,** 13298–13307 (1991).
32. Desjarlais, M. P., Kress, J. D. & Collins, L. A. Electrical conductivity for warm, dense aluminum plasmas and liquids. *Phys. Rev. E* **66,** 025401(R) (2002).
33. Alfè, D. Ab initio molecular dynamics, a simple algorithm for charge extrapolation. *Comput. Phys. Commun.* **118,** 31–33 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements D.G. is supported by CSEDI grant EAR1065597 from the National Science Foundation. C.D. is supported by a Natural Environment Research Council personal fellowship, NE/H01571X/1. M.P. is supported by NERC grant NE/H02462X/1 to D.A. Calculations were performed on the UK national facility HECToR.

Author Contributions D.A. and D.G. designed the project. M.P. and D.A. performed the first principles calculations. C.D. and D.G. performed the thermal history and core stratification calculations. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.A. (d.alfe@ucl.ac.uk).

METHODS

First principles calculations. Calculations were performed using DFT with the same technical parameters used in refs 6, 12–14. We used the VASP code[28], PAW potentials[29,30] with $4s^1 3d^7$ valence configuration, the Perdew–Wang[31] functional, a plane wave cut-off of 293 eV, and single particle orbitals were occupied according to Fermi–Dirac statistics. We tested the effect on the conductivity of the inclusion in valence of semi-core $3s$ and $3p$ states; we found that, as in the zero pressure case[14], this effect is completely negligible.

The electrical conductivity and the electrical component of the thermal conductivity have been calculated using the Kubo–Greenwood formula and the Chester–Thellung–Kubo–Greenwood formula as implemented in VASP[32]. Because of the low mass of the electrons compared to the ions, the conductivities may be calculated by assuming frozen ionic configurations, and averaging over a sufficiently large set representing the typical distribution of the ions at the pressures and temperatures of interest.

Molecular dynamics simulations were performed in the canonical ensemble using cubic simulation cells with 157 atoms and the Γ point, a time step of 1 fs, and an efficient extrapolation of the charge density which speeds up the simulations by roughly a factor of two (ref. 33). Each state point was simulated for at least 6 ps, from which we discarded the first picosecond to allow for equilibration and used the last 5 ps to extract 40 configurations separated by 0.125 ps. This time interval is roughly two times longer than the correlation time, and therefore the configurations are statistically independent from each other. Because of the high temperatures involved, the conductivities converge quickly with respect to **k**-point sampling and size of the simulation cell[14], and we found that with a 157-atom cells and the single **k**-point (1/4,1/4/,1/4) the results are converged to better than 1%.

The ionic component of the thermal conductivity was calculated using the Green–Kubo formula.

Power estimates for the geodynamo. Estimates of the power required to drive the geodynamo are obtained by considering the slow evolution of the Earth using equations describing the balances of energy and entropy in the core. A detailed derivation of these equations can be found in a previous study[5]. Conservation of energy simply equates the heat crossing the CMB to the sources within: specific heat of cooling Q_s, latent heat of freezing Q_L, radiogenic heating Q_r, gravitational energy loss Q_g that is converted into heat by the frictional processes associated with the convection (almost entirely magnetic), and smaller terms[5] involving pressure changes and chemistry that we shall ignore:

$$Q_{\mathrm{CMB}} = Q_s + Q_L + Q_g + Q_r \tag{1}$$

All terms on the right-hand side of equation (1) can be written in terms of either the cooling rate at the CMB, $\mathrm{d}T_0/\mathrm{d}t$, or the amount of radiogenic heating, h. There is no dependence on the conductivities or the magnetic field, which are merely agents by which energy is converted to heat within the core.

These quantities do enter the entropy balance, however. This equation has dissipation terms from thermal and electrical conduction, plus viscosity and molecular diffusion. They are all positive because of the second law of thermodynamics. They are balanced by entropies associated with the power driving the convection: heat pumped in at a higher temperature and removed at a lower temperature (T_{CMB}) and gravitational energy that directly stirs the core and is converted to heat by frictional processes, the heat then being convected and conducted away. Note that entropy from heat is multiplied by a Carnot-like 'efficiency factor', $1/T_{\mathrm{out}} - 1/T_{\mathrm{in}}$ (latent heat is the most efficient because it is released at the highest temperature and removed at the lowest), while the gravitational energy is not, $E_g = Q_g/T_{\mathrm{ICB}}$. Gravitational energy is more efficient at removing entropy and therefore more efficient than heat at generating magnetic field.

$$E = E_s + E_L + E_r + E_g = E_k + E_\sigma + E_\alpha \tag{2}$$

where the four terms on the left-hand side of the second equality represent secular cooling, latent heat release, radiogenic heating and gravitational energy loss. Adiabatic conduction entropy, E_k, is easily estimated from the thermal conductivity and adiabatic gradient and is large, of order $10^8\,\mathrm{W\,K^{-1}}$. The new estimate of conductivity doubles older ones and the higher ICB temperatures increase it still further. Barodiffusion, E_α, is the tendency for light elements to migrate down a pressure gradient and its associated entropy is significant but small, not exceeding $2.5\,\mathrm{MW\,K^{-1}}$ in any of our estimates. Diffusional processes associated with convection and the geodynamo also produce entropy, denoted E_σ, mainly in the small scales. This presents a problem in estimation because the dominant contribution comes from magnetic fields, fluid flows, temperature and compositional fluctuations that cannot be observed and, in many cases, cannot even be simulated numerically. A low value of the power required to drive the dynamo, 0.5 TW (ref. 34), was obtained from a numerical dynamo simulation[35], which at an average temperature of 5,000 K translates into $E_\sigma = 10^7\,\mathrm{W\,K^{-1}}$, an order of magnitude lower than E_k, but the numerical simulation necessarily reduces small scale magnetic fields and the value for the Earth could be much larger. It may well be that future numerical simulations with higher resolution will have higher ohmic dissipation approaching E_k. Magnetic diffusivity is much larger than any other diffusivity in the core, by many orders of magnitude, and in numerical simulations the viscosity, thermal, and molecular diffusivities are replaced with turbulent values to account for unresolved, turbulent, small scale fields. Even so, the associated entropies remain much smaller than those associated with magnetic fields: they are generally ignored, although we should bear in mind that they are all positive and could make a contribution.

Parameter values used to calculate thermal contributions to the energy and entropy balances equations (1) and (2) are taken from Table 1 of a previous study[36], except for the thermal conductivity and the temperatures of the CMB and ICB, which are taken from the present study. Latent heat, Q_L, depends on τ, the difference between the melting and adiabatic gradients at the ICB; the value for the former is taken to be $9\,\mathrm{K\,GPa^{-1}}$ (ref. 36), while the value of the latter is calculated from Fig. 1 of this study. Parameter values used to calculate compositional terms differ slightly from previous work[5], owing to their use of different concentrations for the light elements O, Si and S in the outer core. Concentration enters the calculation of gravitational energy through equation (9) of ref. 5, which, along with equation (8) of ref. 5, is used to define Q_g in equation (18) of ref. 5. Note also Q_g depends on τ. The remaining changes affect the barodiffusion, E_α, which makes a small contribution to the entropy budget (2); for completeness we list the new parameter values required to determine E_α in Supplementary Tables 1 and 2.

Estimating stable layer thicknesses. Radial profiles of the thermal and compositional energy sources that power the dynamo are determined using the equations of a previous study[20], which are derived from the energy balance appropriate for the outer core[5]. The radial profiles represent conductive solutions that satisfy the total CMB heat-flux boundary condition for the temperature, zero CMB mass flux of light elements, and fixed temperature and light element concentration at the ICB[20]. Superimposed on this basic state are the small fluctuations associated with core convection and the dynamo process.

These radial profiles apply to a Boussinesq fluid and hence neglect compressibility effects other than when they act to modify gravity. This necessitates the use of an approximate form for the adiabatic temperature, a simple choice being a quadratic equation expressed in terms of the ICB and CMB temperatures[19]. Despite these simplifications, the CMB heat fluxes computed from equations (23)–(27) of the incompressible model[20] are in good agreement with those obtained from the original equations[5] (see Supplementary Table 3), while the quadratic approximation for the adiabat differs by at most 10 K from the full calculation shown in Fig. 1.

Compositional buoyancy is at least as important for driving the geodynamo as thermal buoyancy (see, for example, ref. 5) and so we require a means of comparing the two in radial profiles, which is readily achieved by multiplying the former by the ratio of compositional and thermal expansion coefficients, α_c/α_T. This simple device converts compositional effects into equivalent thermal effects, thereby allowing all sources of buoyancy to be combined; it is also related to the condition of neutral stability discussed below. (However, it must be understood that the compositional term resulting from this transformation has nothing to do with the gravitational energy, Q_g, which is neglected in the Boussinesq equations[37].) We use the common approach (see, for example, ref. 37) of defining all fluxes that represent sources of buoyancy associated with the convection in terms of a turbulent diffusivity, which is assumed constant. By contrast, the heat flux due to conduction down the adiabatic gradient and the equivalent thermal flux due to barodiffusion must be defined in terms of molecular quantities.

The depth variation of the molecular thermal conductivity obtained from the DFT results is readily incorporated into the formulation of previous work[20]. We write $k = k(r)$ to express the radial variation of the molecular thermal conductivity; equation (8) from ref. 20 must then be replaced by $q_a = \nabla\cdot(k(r)\nabla T_a)$, where ∇T_a is calculated from equation (12) in ref. 20. $k(r)$ is well-approximated by a parabolic conductivity variation, $k(r) = ar^2 + br + c$, which we use to calculate the heat flux down the adiabatic gradient.

To investigate the presence of a stable layer, we use temperature gradients instead of heat fluxes, which are calculated using equations (30)–(34) of a previous study[20] with $k(r)$ replacing k in the numerator of equation (30) of ref. 20. The parameter values are the same as those used to estimate power requirements above. We define the base of the stable layer to be the point of neutral stability as given by Schwarzchild's criterion[38]:

$$\left(\frac{\mathrm{d}T}{\mathrm{d}r}-\frac{\mathrm{d}T_{\mathrm{a}}}{\mathrm{d}r}\right)+\frac{\alpha_{\mathrm{c}}}{\alpha_{\mathrm{T}}}\left(\frac{\mathrm{d}c}{\mathrm{d}r}\right)=0$$

where $\mathrm{d}T/\mathrm{d}r$ is the total temperature gradient, $\mathrm{d}T_{\mathrm{a}}/\mathrm{d}r$ is the adiabatic temperature gradient and $\mathrm{d}c/\mathrm{d}r$ is the total compositional gradient. We write this condition as $T' = T'_{\mathrm{L}} + T'_{\mathrm{s}} + T'_{\mathrm{c}} + T'_{\mathrm{r}} - T'_{\mathrm{a}} = 0$, where the terms represent respectively latent heat, secular cooling, compositional buoyancy, radiogenic heating and the adiabat, and prime indicates differentiation with respect to r (the barodiffusive contribution to $\mathrm{d}c/\mathrm{d}r$ is very small and has been omitted). Possible deviations from the layer thicknesses we obtain using this definition can only be obtained by solving the complete dynamo equations with correct parameters for the Earth, which is impossible at present. We believe this to be the best definition of the base of the layer given the nature of our thermodynamic model.

34. Buffett, B. A. Estimates of heat flow in the deep mantle based on the power requirements for the geodynamo. *Geophys. Res. Lett.* **29,** 1566–1569 (2002).
35. Kuang, W. & Bloxham, J. An Earth-like numerical dynamo model. *Nature* **389,** 371–374 (1997).
36. Gubbins, D., Alfè, D., Masters, T. G., Price, D. & Gillan, M. J. Can the Earth's dynamo run on heat alone? *Geophys. J. Int.* **155,** 609–622 (2003).
37. Anufriev, A. P., Jones, C. A. & Soward, A. M. The Boussinesq and anelastic liquid approximations for convection in the Earth's core. *Phys. Earth Planet. Inter.* **152,** 163–190 (2005).
38. Gubbins, D. & Roberts, P. H. in *Geomagnetism* (ed. Jacobs, J. A.) 30–32 (Academic, 1987).

doi:10.1038/nature11056

Extended leaf phenology and the autumn niche in deciduous forest invasions

Jason D. Fridley[1]

The phenology of growth in temperate deciduous forests, including the timing of leaf emergence and senescence, has strong control over ecosystem properties such as productivity[1,2] and nutrient cycling[3,4], and has an important role in the carbon economy of understory plants[5–7]. Extended leaf phenology, whereby understory species assimilate carbon in early spring before canopy closure or in late autumn after canopy fall, has been identified as a key feature of many forest species invasions[5,8–10], but it remains unclear whether there are systematic differences in the growth phenology of native and invasive forest species[11] or whether invaders are more responsive to warming trends that have lengthened the duration of spring or autumn growth[12]. Here, in a 3-year monitoring study of 43 native and 30 non-native shrub and liana species common to deciduous forests in the eastern United States, I show that extended autumn leaf phenology is a common attribute of eastern US forest invasions, where non-native species are extending the autumn growing season by an average of 4 weeks compared with natives. In contrast, there was no consistent evidence that non-natives as a group show earlier spring growth phenology, and non-natives were not better able to track interannual variation in spring temperatures. Seasonal leaf production and photosynthetic data suggest that most non-native species capture a significant proportion of their annual carbon assimilate after canopy leaf fall, a behaviour that was virtually absent in natives and consistent across five phylogenetic groups. Pronounced differences in how native and non-native understory species use pre- and post-canopy environments suggest eastern US invaders are driving a seasonal redistribution of forest productivity that may rival climate change in its impact on forest processes.

Phenological studies of understory leaf display and gas exchange for native woody species in eastern US (EUS) forests demonstrate a critical period of carbon gain in spring before canopy closure[6,7,13,14]. Significant carbon gain after canopy leaf fall, however, seems to be rare for native forest shrubs[13,14], presumably because lower autumnal solar radiation means deciduous species have less to gain by delaying senescence[1,6]. On the other hand, comparative studies of co-occurring native and non-native understory species[5,8,9] have demonstrated both earlier and later carbon gain for non-natives, suggesting that extended leaf phenology could be an important mechanism of invader establishment in EUS forests[12,15]. To determine whether this pattern is general across a broad taxonomic sample of native and non-native species, I established an experimental garden of three replicate blocks of 73 woody species common to EUS deciduous forests, including native and non-native representatives of several phylogenetic groups (*Celastrus*, *Euonymus* in Celastraceae; *Elaeagnus* in Elaeagnaceae; *Frangula*, *Rhamnus* in Rhamnaceae; *Lonicera* in Caprifoliaceae; *Viburnum* in Adoxaceae) and other unrelated but widespread natives (Supplementary Table 1). All 30 non-native species are naturalized in the EUS and all but eight are currently managed as forest invaders[16]. From the onset of local forest canopy closure (approximately 20 May) until canopy leaf fall (approximately 24 Oct), plants were grown under 80% shade to simulate a deciduous understory light environment. For three growing seasons (2008–2010), we monitored the timing of spring foliar bud and leaf development, biweekly leaf production and chlorophyll (Chl) content, and monthly photosynthetic rate on select leaves at a range of light levels (50–800 μmol pm^{-2} s^{-1}). Although not all species were measured each year, a similar number of native and non-native species were monitored annually and data sets for most species involved at least 2 years (Supplementary Table 1).

The timing of leaf emergence across species was sensitive to interannual variation in spring temperatures (Supplementary Fig. 1), with all stages of bud development occurring several weeks earlier in the warmer spring of 2010 ($P < 0.001$, year contrasts in Mann–Whitney *U*-tests adjusted for multiple testing; Fig. 1). However, the timing of early stages of bud activity was not significantly different between native and non-native species for any year (Fig. 1). As a group, non-native species showed earlier budburst in 2010 and earlier full extension of true leaves in 2008 and 2010, but differences in median date were small (3, 1 and 2 days, respectively) and there was no evidence that non-natives were more responsive to the warmer spring of 2010 (year–nativity interaction, $P > 0.5$). In contrast, the timing of autumn leaf fall for non-native species as a group was delayed by as much as 28 days compared with natives (Fig. 1). The median date of 50% leaf fall for native species across 2008–2010 was 16 September, and for non-native species 13 October; the same comparison for 90% leaf fall was 27 October and 9 November. Autumn leaf phenology did not vary significantly across years (Akaike information criterion of models with and without year random effect, 2,512 versus 2,495, $P < 0.001$, likelihood ratio = 15.58 on 1 d.f.). Owing to the large difference in autumn phenology, non-native species on average had a growing season 29 days longer than natives, using days of extant true leaves until 50% leaf fall, which amounts to an extension over the native growing season of 19%.

Leaf Chl and gas exchange measurements supported strong maintenance of leaf function for non-native species after leaves of most natives had senesced. Non-native species retained high levels of leaf Chl in autumn compared with native species, despite similar spring Chl phenology (Fig. 2). In spring and summer, Chl content of both groups was determined by whether leaves were produced in sun or shade, with sun leaves peaking in Chl content in mid-August, 16 days before shade leaves ($P < 0.001$, $t = -3.76$ on 510 d.f.). In contrast, Chl content was identical for sun and shade leaves after mid September within both non-native and native groups, but non-native species showed significant delays in Chl breakdown, with an average difference in the date of 50% peak Chl loss between natives and non-natives of more than 2 weeks (Fig. 2; $P < 0.001$, $t = -3.54$ on 83 d.f.). Photosynthetic rates for both high (800 photosynthetic photon flux density (PPFD)) and low (100 PPFD) light levels followed the seasonal trajectory of Chl content for both native and non-native species, with peaks in summer, relatively high rates in spring, and no differences between groups (Supplementary Fig. 2). In autumn, photosynthetic rates for most native species (30 out of 43 species) declined to zero because of loss of live leaves, whereas most non-native species (21 out of 30 species) continued to assimilate carbon. Of the subset of natives and non-natives with live leaves after shade cloth removal, non-natives

[1]Department of Biology, Syracuse University, 107 College Place, Syracuse, New York 13244, USA.

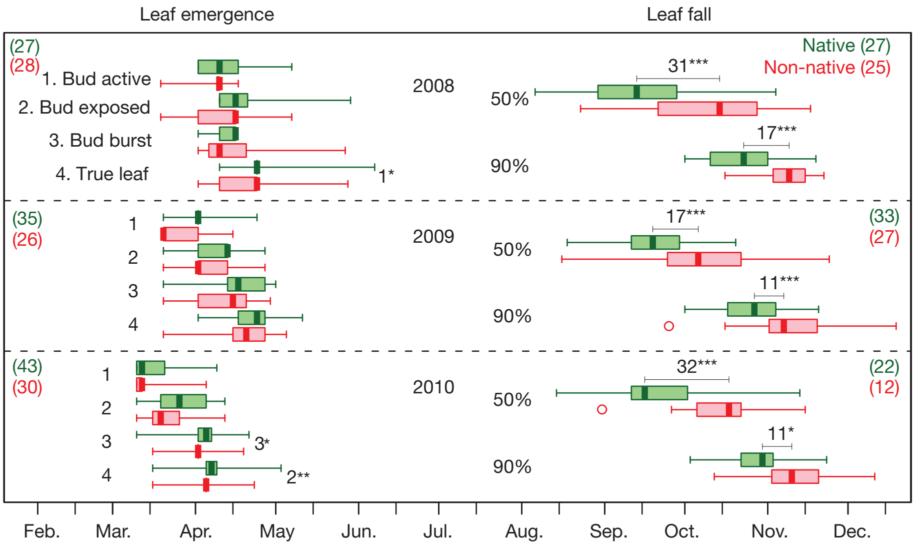


Figure 1 | Seasonal patterns of leaf emergence and leaf fall for native and non-native species over three growing seasons. Boxplots show data range with boxed first and third quartiles, median as heavy line, and point outliers; numbers of species for each group are indicated. Leaf emergence was monitored at 2- to 5-day intervals using a classification of budbreak stages and dates at which 50% and 90% of total leaves had fallen were interpolated from biweekly monitoring. Values indicate median difference in days between natives (green) and non-natives (red). *P* values are from Mann–Whitney *U*-tests adjusted for multiple testing (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$).

had marginally significant higher rates of photosynthesis in high light (Supplementary Fig. 2).

To quantify the impact of extended growth phenology on the total annual carbon (C) gain of native and non-native species, I estimated daily C assimilation for each species with a stochastic simulation model using empirical distributions of seasonal leaf production and photosynthetic capacity (see Methods). Most species captured a significant portion of their annual C assimilate before canopy closure on 20 May (Fig. 3a), with two-thirds of the species getting at least 10% of their annual C in the spring (mean = 14%). Natives as a group had a larger contribution of pre-canopy C gain than non-natives, both overall ($P < 0.05$, $t = 2.06$ on 70 d.f.) and after accounting for phylogenetic groups with a random effect ($P < 0.05$, $F = 4.22$ on 1, 63 d.f.), although separate tests within groups were not significant (Fig. 3b). Of those species of highest spring C gain, only one non-native species (*Lonicera* × *bella*) made the top 10 and only four made the top 20. In contrast, post-canopy C assimilation (after 24 October) was strongly biased towards non-native species and virtually absent in natives, with only three native species (*Diervilla rivularis*, *Diervilla lonicera*, *Lonicera sempervirens*) gaining more than 10% of their annual C in autumn, and more than half of the natives (25 out of 43 species) obtaining less than 1% (Fig. 3c). Nearly half of the non-native species (13 out of 30 species) obtained at least 5% of their C in autumn, and seven gained more than 10%, up to a maximum of 21% (*Lonicera fragrantissima*). The strong non-native C gain advantage in autumn was highly significant when controlling for phylogenetic group as a

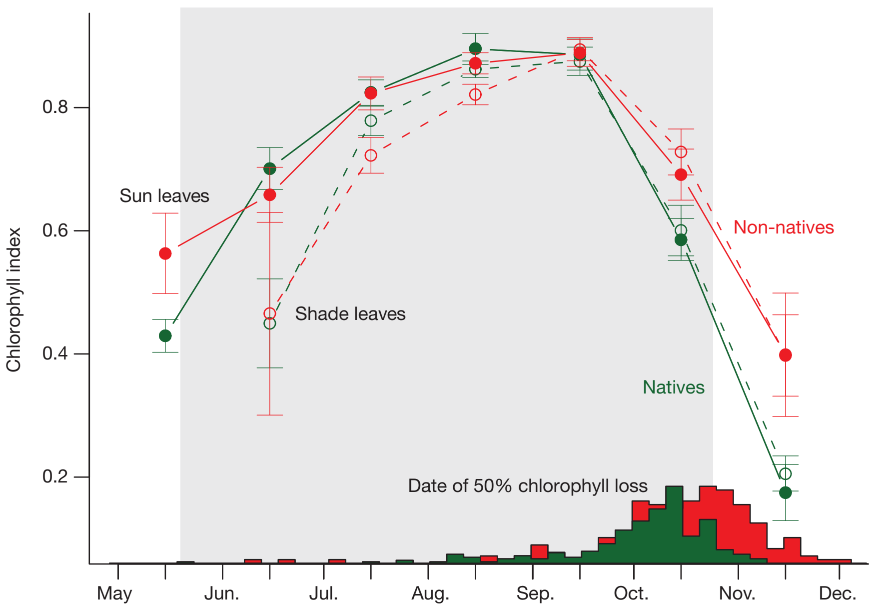


Figure 2 | Relative leaf Chl content for native and non-native species. Mean (± s.e.m.) content for native (green) and non-native (red) species are grouped by whether leaves were produced before (filled circles and lines) or after (open circles and dashed lines) canopy shading (grey region). Histograms show distributions of the date of 50% Chl loss, relative to peak Chl reading per leaf, for native ($n = 354$ leaves) and non-native species ($n = 253$) pooled across 2009 and 2010 growing seasons.

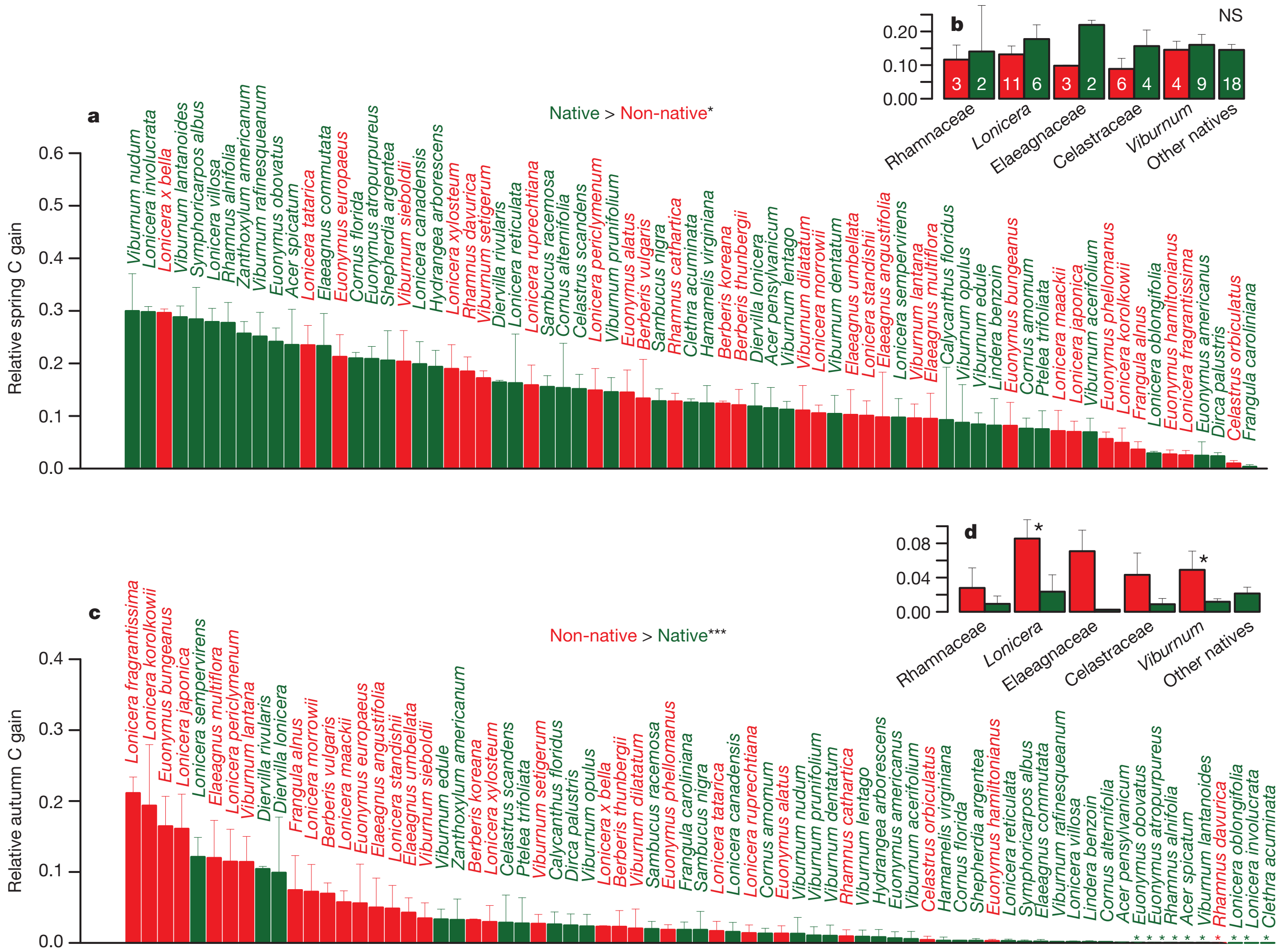


Figure 3 | Proportion of total annual C assimilated in spring and autumn for native and non-native species. Values of assimilation before (**a**, **b**) and after (**c**, **d**) shade cloth placement for native (green) and non-native (red) species were estimated by stochastic simulation of daily leaf area, light levels and photosynthetic capacity using empirical measurements (2008–2010). Values are means (± s.e.m.) of 1,000 permutations incorporating measured variation in individual and interannual leaf production and photosynthetic light curves. Inset figures (**b**, **d**) show mean (± s.e.m.) species values summarized by phylogenetic group, with sample sizes indicated. Black asterisks indicate statistical significance for overall native–non-native comparisons and Mann–Whitney *U*-tests within groups (*$P < 0.05$, ***$P < 0.001$). Coloured asterisks denote autumn C gain less than 0.5%. NS, not significant.

random effect ($P < 0.001$, $F = 15.81$ on 1, 63 d.f.), although small sample sizes within groups other than *Lonicera* and *Viburnum* precluded detection of significant nativity trends when groups were analysed separately (Fig. 3d).

Although non-native species seem to inhabit an autumn niche that is rare in the native woody flora, it is not clear from these data whether this constitutes their primary fitness advantage over natives. Across all species, total annual C gain was significantly associated with both early and late growth phenology ($P < 0.05$, $F = 5.28$ and 11.14 on 1, 59 d.f. for spring and autumn relative C gain, respectively, including a random phylogenetic group effect), confirming that earlier budbreak and delayed leaf senescence behaviours contribute to annual growth rates. However, non-native species overall did not assimilate more C annually than natives ($P > 0.3$, $F = 0.86$ on 1, 63 d.f.), partly because of the significant spring advantage of natives. Because spring phenology was variable but autumn phenology was not, the relative contribution of extended autumn phenology to the success of invaders may depend on spring temperatures, and it is possible that earlier springs will favour native species. This analysis does not include seasonal C losses from night-time respiration, however, which for many deciduous species are highest in spring[13].

The presumed costs of late leaf display for winter deciduous species are nutrient loss from lack of resorption in frost-damaged leaves[17] and shoot damage from delayed tissue hardening[18]. Why should colonizing deciduous species from temperate Eurasian environments not bear these costs? One possibility is that non-natives are better adapted to the warmer autumn temperatures experienced in the EUS over the past several decades[19], or are more responsive to elevated levels of soil nitrogen availability from industrial pollution[20], reducing their need to maximize resorption. Recent environmental changes cannot be a general explanation, however, given that many of the non-native species in the present study have been invasive for over a century[21]. On the other hand, increased soil nutrient fluxes in North American forests from Eurasian earthworm invasion may have been coincident with plant introductions[22]; it is conceivable that non-native species, having co-evolved with earthworms, evolved a nutrient-use strategy that is less dependent on autumn resorption, thus explaining observed associations of invasive shrub and earthworm abundance[23]. It is also notable that most of the invasive shrubs and lianas in EUS are from East Asia[16], a region that experienced significantly less climate disruption than EUS during the Pleistocene[24]. It is possible that the more restricted growth phenology of the EUS flora today is a relictual behaviour from shorter Pleistocene growing seasons[25], leaving some East Asian species 'pre-adapted' to the modern EUS forest environment[26]. However, the extended growth phenology of several species from Europe, a region of more severe Pleistocene climate disruption than EUS, would still require explanation.

Although it is not possible from this study to quantify the ecosystem impacts of extended foliar phenology of understory non-native species under natural forest conditions, eddy flux data[1,27,28] indicate that even minor changes in growing season duration can have a significant effect on forest productivity. In this context the impacts of forest invaders extending the period of C assimilation into autumn by several weeks may rival that of climate forcing[1], although additional studies of net C balance are needed to quantify potential differences in the seasonal C contribution of native and non-native species[28]. The higher autumnal activity of invaders may also lead to significant shifts in nutrient cycling, particularly if leaf nitrogen resorption is reduced in non-natives as a result of delayed senescence, causing significant changes in forest-floor litter quality[29]. Although their contribution to forest standing biomass is small, understory species can have disproportionate impacts on ecosystem fluxes[3,30], which suggests the extended understory growing season in deciduous forests resulting from continuing invasions by non-native shrubs, lianas and some herbs[10] may be a major driver of anthropogenic ecosystem change in eastern North America.

METHODS SUMMARY

The garden was established in 2006–2007 in Syracuse, New York, USA (43° 03′ N, 76° 09′ W). It included genera for which there exist at least one native species and one non-native invasive species present in EUS forests (see ref. 16 for details on 'non-native', 'invasive' and habitat criteria), in addition to 16 widespread but unrelated native species (Supplementary Table 1). Transplants were collected from the wild in central New York where possible (20 native and 9 alien species); those absent from wildlands in the region were obtained from commercial growers of similar latitude (Supplementary Table 1). Plants were spaced 1 m apart in three replicate blocks and covered with 80% knitted black shade cloth from 20 May to 24 October each monitoring year (2008–2010). We monitored spring bud and leaf development on each plant by photographing select nodes at 2- to 5-day intervals from early March to mid-May and classifying images according to five development stages. Leaf demography and Chl content (CCM-200 sensor) were monitored at 2-week intervals on five branches selected at random per individual per year. Photosynthesis was monitored monthly for each individual at intensities of 800, 300, 100 and 50 μmol photon $m^{-2} s^{-1}$ (LI-COR 6400; 400 μmol CO_2 mol^{-1}, 700 μmol s^{-1} flow rate, 20 °C). Fitted parameters for the non-rectangular hyperbolic light curve function (apparent quantum yield (AQY), A_m, R_d, α) were used in seasonal C gain simulations, incorporating daily leaf area from leaf demography data and daily light levels from a photosynthetically active radiation sensor. C gain summaries were derived from 1,000 annual permutations; for each, photosynthetic parameters were randomly sampled from species- and season-specific normal distributions and daily leaf area was determined by random samples from empirical distributions of daily leaf counts.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 15 December 2011; accepted 15 March 2012.
Published online 25 April 2012.

1. Richardson, A. D. *et al.* Influence of spring and autumn phenological transitions on forest ecosystem productivity. *Phil. Trans. R. Soc. B* **365,** 3227–3246 (2010).
2. Polgar, C. A. & Primack, R. B. Leaf-out phenology of temperate woody plants: from trees to ecosystems. *New Phytol.* **191,** 926–941 (2011).
3. Muller, R. N. & Bormann, F. H. Role of *Erythronium americanum* Ker. in energy flow and nutrient dynamics of a northern hardwood forest ecosystem. *Science* **193,** 1126–1128 (1976).
4. Ehrenfeld, J. G. Effects of exotic plant invasions on soil nutrient cycling processes. *Ecosystems* **6,** 503–523 (2003).
5. Harrington, R. A., Brown, B. J. & Reich, P. B. Ecophysiology of exotic and native shrubs in southern Wisconsin. I. Relationship of leaf characteristics, resource availability, and phenology to seasonal patterns of carbon gain. *Oecologia* **80,** 356–367 (1989).
6. Augspurger, C. K., Chesseman, J. M. & Salk, C. F. Light gains and physiological capacity of understorey woody plants during phenological avoidance of canopy shade. *Funct. Ecol.* **19,** 537–546 (2005).
7. Rothstein, D. E. & Zak, D. R. Photosynthetic adaptation and acclimation to exploit seasonal periods of direct irradiance in three temperate, deciduous-forest herbs. *Funct. Ecol.* **15,** 722–731 (2001).
8. Schierenbeck, K. A. & Marshall, J. D. Seasonal and diurnal patterns of photosynthetic gas exchange for *Lonicera sempervirens* and *L. japonica* (Caprifoliaceae). *Am. J. Bot.* **80,** 1292–1299 (1993).
9. Xu, C.-Y., Griffin, K. L. & Schuster, W. S. F. Leaf phenology and seasonal variation of photosynthesis of invasive *Berberis thunbergii* (Japanese barberry) and two co-occurring native understory shrubs in a northeastern United States deciduous forest. *Oecologia* **154,** 11–21 (2007).
10. Myers, C. V. & Anderson, R. C. Seasonal variation in photosynthetic rates influences success of an invasive plant, garlic mustard (*Alliaria petiolata*). *Am. Midl. Nat.* **150,** 231–245 (2003).
11. Wolkovitch, E. M. & Cleland, E. E. The phenology of plant invasions: a community ecology perspective. *Front. Ecol. Environ* **9,** 287–294 (2011).
12. Willis, C. G., Ruhfel, B. R., Primack, R. B., Miller-Rushing, A. J. & Losos, J. B. Favorable climate change response explains non-native species' success in Thoreau's Woods. *PLoS ONE* **5,** e8878 (2010).
13. Gill, D. S., Amthor, J. S. & Bormann, F. H. Leaf phenology, photosynthesis, and the persistence of saplings and shrubs in a mature northern hardwood forest. *Tree Physiol.* **18,** 281–289 (1998).
14. Augspurger, C. K. & Bartlett, E. A. Differences in leaf phenology between juvenile and adult trees in a temperate deciduous forest. *Tree Physiol.* **23,** 517–525 (2003).
15. Webster, C. R., Jenkins, M. A. & Jose, S. Woody invaders and the challenges they pose to forest ecosystems in the Eastern United States. *J. For.* **104,** 366–374 (2006).
16. Fridley, J. D. Of Asian forests and European fields: Eastern U.S. plant invasions in a global floristic context. *PLoS ONE* **3,** e3630 (2008).
17. May, J. D. & Killingbeck, K. T. Effects of preventing nutrient resorption on plant fitness and foliar nutrient dynamics. *Ecology* **73,** 1868–1878 (1992).
18. Saxe, H., Cannell, G. R., Johnsen, Ø., Ryan, M. G. & Vourlitis, G. Tree and forest functioning in response to global warming. *New Phytol.* **149,** 369–400 (2001).
19. Mitchell, T. D. & Jones, P. D. An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol.* **25,** 693–712 (2005).
20. Aber, J. D., Nadelhoffer, K. J., Steudler, P. & Melillo, J. M. Nitrogen saturation in northern forest ecosystems. *Bioscience* **39,** 378–387 (1989).
21. Mack, R. N. Plant naturalizations and invasions in the Eastern United States: 1634–1860. *Ann. Mo. Bot. Gard.* **90,** 77–90 (2003).
22. Frelich, L. E. *et al.* Earthworm invasion into previously earthworm-free temperate and boreal forests. *Biol. Invasions* **8,** 1235–1245 (2006).
23. Nuzzo, V. A., Maerz, J. C. & Blossey, B. Earthworm invasion as the driving force behind plant invasion and community change in Northeastern North American forests. *Conserv. Biol.* **23,** 966–974 (2009).
24. Ehlers, J. & Gibbard, P. L. The extent and chronology of Cenozoic global glaciation. *Quat. Int.* **164,** 6–20 (2007).
25. Lechowicz, M. J. Why do temperate deciduous trees leaf out at different times? Adaptation and ecology of forest communities. *Am. Nat.* **124,** 821–842 (1984).
26. Mack, R. N. Phylogenetic constraint, absent life forms, and preadapted alien plants: a prescription for biological invasions. *Int. J. Plant Sci.* **164,** S185–S196 (2003).
27. Goulden, M. L., Munger, J. W., Fan, S.-M., Daube, B. C. & Wofsy, S. C. Exchange of carbon dioxide by a deciduous forest: response to interannual climate variability. *Science* **271,** 1576–1578 (1996).
28. Piao, S. *et al.* Net carbon dioxide losses of northern ecosystems in response to autumn warming. *Nature* **451,** 49–52 (2008).
29. Liao, C. *et al.* Altered ecosystem carbon and nitrogen cycles by plant invasion: a meta-analysis. *New Phytol.* **177,** 706–714 (2008).
30. Chapin, F. S. Nitrogen and phosphorus nutrition and nutrient cycling by evergreen and deciduous understory shrubs in an Alaskan black spruce forest. *Can. J. For. Res.* **13,** 773–781 (1983).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements I acknowledge technical assistance and advice from A. Craddock, S. Anderson, E. Fridley, S. Kelley, A. Surace, L. Giencke and D. Frank, and project support from Syracuse University.

Author Contributions J.D.F. designed the study, supervised data collection, performed the analyses and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.D.F. (fridley@syr.edu).

METHODS

Study design and focal species. In 2006–2007 I established an experimental shade garden in Syracuse, New York, USA (43°03′ N, 76°09′ W), including three replicate blocks of 73 deciduous shrub and liana species (Supplementary Table 1). Species included congeners of eight genera (*Berberis*, *Celastrus*, *Elaeagnus*, *Euonymus*, *Frangula*, *Lonicera*, *Rhamnus*, *Viburnum*) for which there exists at least one species native to forests of the EUS and one non-native invader present in EUS forests or woodlands (see ref. 16 for details on habitat designations and 'non-native' and 'invader' criteria). A ninth species group consisted of common native EUS forest shrubs lacking EUS non-native invasive congeners. Transplants of most individuals were planted in 2006–2007, although a few were obtained in subsequent years; individuals were not monitored the year they were transplanted. Transplants were collected from the wild in central New York where possible (20 native and 9 non-native species); those absent from wildlands in the region were obtained from commercial growers of roughly the same latitude (including Forestfarm Nursery, Williams, Oregon, and Musser Forests, Indiana, Pennsylvania; seven species could be found only from southern US sources, including four natives and three non-natives; Supplementary Table 1). Individuals were spaced approximately 1 m apart beneath a wooden frame structure supporting 80% knitted black polypropylene shade cloth (DeWitt), deployed seasonally to coincide with local dates of forest canopy closure (approximately 20 May) and canopy leaf fall (approximately 24 October). This light regime approximates deciduous woodland conditions (midday PPFD measurements on a clear day under the shade cloth near the summer solstice peaked around 350 μmol m^{-2} s^{-1}, well above understory light levels in a mature, stratified temperate deciduous forest[31]) and was chosen as a compromise between light levels too high for forest conditions and too low to promote significant short-term growth. Estimates of percentage carbon gain before and after shade cloth placement are therefore conservative for a heavy canopy, where summer photosynthetic gain is expected to be lower than that observed in this study. For those species whose growth phenology has been examined elsewhere at a similar latitude, the behaviours displayed in the Syracuse garden were similar (including *Lonicera* × *bella* and *Rhamnus cathartica*[5], *Viburnum lantanoides*[13], *Lindera benzoin*[6] and *Berberis thunbergii*[9]), suggesting the patterns reported here are widely applicable to north-temperate latitudes. Neutral shade under polypropylene is also enriched in red:far-red ratio compared with natural forest understories; however, shade-tolerant plants like those used in the present study have been shown to be relatively insensitive to red:far-red ratio[32]. In several cases species-level replication was reduced by mortality, including ten species in the present study that were represented in only one block (Supplementary Table 1). Because comparisons of the attributes of particular species are not the focus of the present study, these data were retained for overall comparisons of native and non-native groups and statistical results were checked for dependence on the inclusion of one-replicate species. An additional target species, the EUS native *Berberis canadensis*, was excluded from the present study because garden individuals are suspected to be hybrids of *B. canadensis* and *B. thunbergii*.

Leaf budbreak. From 2008 onwards, we monitored spring bud and leaf development on each plant by photographing select nodes at 2- to 5-day intervals from early March to mid May. Buds from each image were classified according to five development stages: (1) dormant; (2) active (apparent bud swelling, scale development, visibility of inner scales, or scales changing in colour); (3) exposed (inner bud tissue apparent, including secondary cataphylls, transitional leaves, or tips of first leaves; first leaf reflexion in species lacking bud scales); (4) broken (general loosening of all bud structures including inner leaves, some exposure of leaf lamina; second leaf reflexion in species lacking bud scales); and (5) flushed/true leaf (full laminar surface of true leaf visible). Distributions of the timing of bud development across years and for native and non-native species were non-normal and compared by Mann–Whitney *U*-tests[33] with *P* values adjusted for multiple testing across years[34].

Leaf demography and senescence. We selected five healthy terminal branches at random on each individual (maximum 15 branches per species per year) before budbreak in early spring to monitor leaf production and senescence at 2 week intervals, with initial leaf emergence monitored at 3-day intervals to capture rapid spring development. Total extant leaves on existing buds and new shoots on each branch were counted at each interval. A new leaf was counted once it had reflexed by 20°, and leaf 'death' was defined as greater than 50% chlorosis. Branches damaged by herbivory or inadvertent breakage were excluded from further analysis. Leaf Chl content for select leaves was measured with a Chl meter (CCM-200, Opti-Sciences) by averaging three to five readings per leaf (avoiding the midrib) from July to December in 2009 and throughout the growing season in 2010. The CCM-200 measures the ratio of radiation transmitted through the leaf at wavelengths of 940 and 660 nm. Tests[35] indicate a correlation of CCM-200 index readings and total leaf Chl of $R^2 > 0.95$. Where possible, leaves were selected that emerged both before and after shade cloth placement. Leaf Chl data were pooled across 2009 and 2010 for analysis. For each monitored leaf, Chl values for standardized dates (15th of each month, May–November) were linearly interpolated from time series data, and species means were relativized separately for sun and shade leaves by their maximum Chl index reading (values between 0 and 1). Means and standard errors were then calculated separately for sun and shade leaves across native and non-native groups. Overall correlates of Chl content were tested in a mixed effects model using all 615 leaves monitored, using 'nativity' and 'sun/shade' as main effects and 'species' and 'year' as random effects[36].

Seasonal photosynthesis. Net leaf photosynthetic rates were monitored on a monthly basis (2008–2010) on a leaf of each individual using photosynthetic light curves (LI-COR 6400 with red–blue light-emitting diode light source, LI-COR Biosciences). We used a customized program (400 μmol CO_2 mol^{-1}, 700 μmol s^{-1} flow rate, 20 °C) that logged at 20-s intervals, starting with equilibration for four minutes at 800 μmol photon m^{-2} s^{-1} and descending to 300, 100 and 50 for 2 min each, which preliminary trials in 2007 suggested was sufficient for leaf equilibration to different light levels. Measurements were taken daily on a species-rotating basis between 9:00 and 12:00 with replicate individuals of each species done on the same day. Mean photosynthetic rate of each species at high (800) and low light (100 PPFD) was modelled for season (before 20 May, 20 May to 24 October, after 24 October, averaged across years) and nativity in a linear mixed-effects model including season and species as random effects[36]. *A priori* multiple comparisons of native-non-native contrasts for each season were tested for significance using the simultaneous testing procedure of ref. 37.

Photosynthesis data were interpolated to hourly estimates over the growing season using the following procedure. For each measurement interval, light curve data were fitted to the four-parameter non-rectangular hyperbolic function[38] describing net CO_2 uptake (A) in units of μmol CO_2 m^{-2} s^{-1}:

$$A=\left\{\mathrm{AQY}\times\mathrm{PPFD}+A_{\mathrm{m}}-\sqrt{(\mathrm{AQY}\times\mathrm{PPFD}+A_m)^2-4\times\alpha\times\mathrm{AQY}\times\mathrm{PPFD}\times A_m}\right\}\Big/2\alpha-R_{\mathrm{d}}$$

where AQY is apparent quantum yield, PPFD is irradiance (μmol m^{-2} s^{-1}), A_m is maximum photosynthetic rate, R_d is dark respiration, and α is a unitless shape parameter describing curve convexity. To take advantage of all 30 values from the above light curve program (10 min of 20-s intervals) to maximize the robustness of model fit, I used nonlinear quantile regression[39] to estimate parameter values by fitting the 95th quantile rather than the mean response, reflecting the tendency of photosynthesis rates to equilibrate gradually to a maximum for each PPFD level. This procedure produced very high goodness-of-fit values as assessed by model residuals and fitted values (median R^2 of 2,100 curves = 0.98). In this way mean and standard errors for AQY, A_m, R_d and α were estimated for all species for each month leaves were present during the years of measurement (note that not all species were monitored in all years; mean $n = 24$ for each parameter per species). These parameter values were then used to model seasonal variation in AQY, A_m and R_d by interpolation using a regression spline generalized additive model[40] relating Julian date to each parameter, with values weighted by their precision (standard $error^{-1}$). Daily estimates of AQY, A_m and R_d (mean and standard error) were used in the carbon assimilation estimates (for species means see Supplementary Table 1). The shape parameter α varied little seasonally and was represented by its overall species mean.

Daily carbon gain estimates. The contribution of net photosynthetic activity in early spring and late autumn to total annual carbon gain was estimated for each species by stochastically simulating potential carbon assimilation over each day of the growing season. Given leaf-level photosynthetic properties and total leaf area of a particular species on Julian day *t*, and a diurnal light regime for day *t* at plant level from light sensor data, the simulation estimated whole-plant potential carbon gain over a 24-h period. The simulation was stochastic because all parameters (except for PPFD) were treated as random variables, with photosynthetic parameters sampled from their estimated daily distributions (see above) and daily leaf area estimates derived from empirical distributions of branch leaf counts. Daily peak PPFD values (2008–2010) from an LI-COR quantum sensor outside the period of shade cloth placement (full sun) were fitted to a daily model describing the linear increase or decrease of PPFD between winter and summer solstices[41], assuming clear-sky conditions. A similar model was fitted to light sensor data under the shade cloth (20 May to 24 October), with a rise in PPFD to the 21 June maximum and a decline (of equal rate) thereafter. The full seasonal PPFD curve is shown in Supplementary Fig. 3. Peak daily PPFD values were then used to interpolate light values to 30 min intervals throughout

the year using local sunrise, sunset and solar noon data, by means of linear regression (0 PPFD at sunrise to peak PPFD at solar noon, then back to 0 at sunset). For each day a plant had functional leaves, it assimilated carbon according to the function:

$$\text{daily C gain (in grams)} = \sum_{i}^{1-48} \text{area} \times L \times f(\text{PPFD}_i, A_{\text{m}}, R_{\text{d}}, \text{AQY}, \alpha) \times 1800 \times (12 \times 10^{-6})$$

where area is the area of a fully expanded leaf, L is the number of extant functional leaves on day t and f is the above non-rectangular hyperbolic function, including PPFD levels for each 30-min interval (i). The constants convert seconds to 30 min intervals and μmol to grams. Photosynthesis parameters AQY, A_{m}, and R_{d} were sampled randomly from a normal distribution of mean and standard error fitted from the above daily generalized linear model interpolation, and area and α were sampled from normal distributions using their overall means and standard errors (leaf areas included 5–10 samples per species). Owing to the large variability in leaf production per branch for most species, L was sampled from the empirical distribution of daily leaf counts across branch–years. The simulation included 1,000 permutations of annual trajectories of daily carbon gain for each species. Two native species, *Lonicera hirsuta* and *Shepherdia canadensis*, were strong positive outliers in the distribution of relative spring C gain (both near 50%), owing to all replicates showing very early leaf fall in 2010 (mid July), possibly the result of insect herbivory (*S. canadensis*) and powdery mildew (*L. hirsuta*). So as not to bias the native distribution of C gain phenology as a result of pest damage, these species were removed from C gain analysis. This had no qualitative effect on native–non-native C gain comparisons.

31. Canham, C. D. *et al.* Light regimes beneath closed canopies and tree-fall gaps in temperate and tropical forests. *Can. J. For. Res.* **20,** 620–631 (1990).
32. Morgan, D. C. & Smith, H. A. systematic relationship between phytochrome-controlled development and species habitat, for plants grown in simulated natural radiation. *Planta* **145,** 253–258 (1979).
33. Hothorn, T. & Hornik, K. exactRankTests: exact distributions for rank and permutation tests. R package version 0.8-19 (2010).
34. Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75,** 800–802 (1988).
35. Richardson, A. D., Duigan, S. P. & Berlyn, G. P. An evaluation of noninvasive methods to estimate foliar chlorophyll content. *New Phytol.* **153,** 185–194 (2002).
36. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & the R Development Core Team. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-102 (2011).
37. Hothorn, T., Bretz, F. & Westfall, P. Simultaneous inference in general parametric models. *Biometrical J.* **50,** 346–363 (2008).
38. Lambers, H., Chapin, F. S. & Pons, T. L. *Plant Physiological Ecology* (Springer, 1998).
39. Koenker, R. quantreg: Quantile Regression. R package version 4.44 (2009).
40. Wood, S. N. Fast stable direct fitting and smoothness selection for generalized additive models. *J. R. Stat. Soc. B* **70,** 495–518 (2008).
41. Hutchison, B. A. & Matt, D. R. The distribution of solar radiation within a deciduous forest. *Ecol. Monogr.* **47,** 185–207 (1977).

LETTER

doi:10.1038/nature10987

Restoration of grasp following paralysis through brain-controlled stimulation of muscles

C. Ethier[1], E. R. Oby[1], M. J. Bauman[2] & L. E. Miller[1,3,4]

Patients with spinal cord injury lack the connections between brain and spinal cord circuits that are essential for voluntary movement. Clinical systems that achieve muscle contraction through functional electrical stimulation (FES) have proven to be effective in allowing patients with tetraplegia to regain control of hand movements and to achieve a greater measure of independence in daily activities[1,2]. In existing clinical systems, the patient uses residual proximal limb movements to trigger pre-programmed stimulation that causes the paralysed muscles to contract, allowing use of one or two basic grasps. Instead, we have developed an FES system in primates that is controlled by recordings made from microelectrodes permanently implanted in the brain. We simulated some of the effects of the paralysis caused by C5 or C6 spinal cord injury[3] by injecting rhesus monkeys with a local anaesthetic to block the median and ulnar nerves at the elbow. Then, using recordings from approximately 100 neurons in the motor cortex, we predicted the intended activity of several of the paralysed muscles, and used these predictions to control the intensity of stimulation of the same muscles. This process essentially bypassed the spinal cord, restoring to the monkeys voluntary control of their paralysed muscles. This achievement is a major advance towards similar restoration of hand function in human patients through brain-controlled FES. We anticipate that in human patients, this neuroprosthesis would allow much more flexible and dexterous use of the hand than is possible with existing FES systems.

Worldwide, over 130,000 people each year survive spinal cord injury (SCI) but sustain extensive paralysis[4]. Approximately half of these injuries occur above the sixth cervical vertebra, thereby affecting all four limbs. Most of these patients indicate that regaining the ability to grasp objects would provide the greatest practical benefit compared to regaining other lost functions[5].

For this reason, considerable effort has been devoted to the development of FES systems to restore voluntary grasp[1,2,6]. These systems rely on residual movement or muscle activity to control electrical activation of hand muscles. Because of the complexity of the necessary patterns of muscle activation, current FES systems produce only one or two grasps using pre-programmed stimulus trains that must be customized for each user[7]. This is effective because many objects can be grasped adequately with only palmar or pinch grasp. However, normal hand use is much more complex than this. Furthermore, using the motion of one body part to control that of another inevitably increases the associated cognitive burden. If FES is to provide hand movements that are close to normal, a more natural control signal of higher dimension than that available through residual motion will be necessary.

Fortunately, the rapid development of the brain machine interface (BMI) provides promising new means by which more flexible and dexterous movements might be controlled. However, despite the initial demonstration of strong force-related discharge in the primary motor cortex (M1)[8], virtually all existing BMIs extract only kinematic information from the brain. This bias is ironic, as the first study to decode signals from simultaneously recorded neurons found that force was more strongly represented than movement in M1 (ref. 9).

Only a small number of groups have pursued the possibility of using kinetic (force-related) information as a real-time control signal for a BMI, through the prediction of grasp force[10,11], joint torque[12] or muscle activity[10,13,14]. We showed previously that despite paralysis produced by peripheral nerve block, monkeys could accurately modulate the magnitude of isometric flexion and extension wrist torque using cortically controlled FES[15,16,17]. Related results were also reported by a group that operantly conditioned monkeys to modulate the activity of one or two individual neurons whose discharge directly controlled stimulation of individual muscles[18].

We performed the current experiments with two monkeys trained to pick up weighted rubber balls and to convey them to an opening at the top of a dispenser (Fig. 1). After training, each monkey was implanted with a multi-electrode recording array in the hand area of M1. In a separate surgical procedure, we implanted intramuscular electrodes for recording and stimulation of hand and forearm muscles. Figure 2a shows the neural discharge recorded under normal conditions from a representative session. Most of these 104 neuronal signals

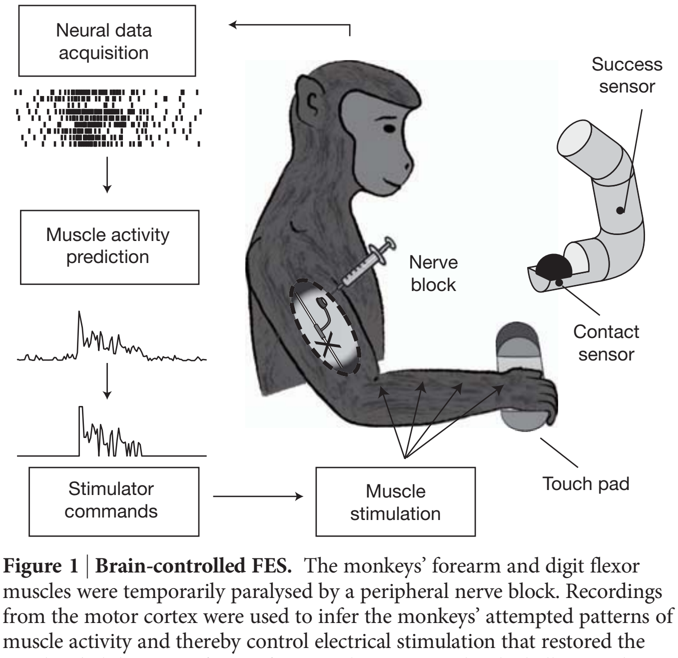


Figure 1 | Brain-controlled FES. The monkeys' forearm and digit flexor muscles were temporarily paralysed by a peripheral nerve block. Recordings from the motor cortex were used to infer the monkeys' attempted patterns of muscle activity and thereby control electrical stimulation that restored the monkeys' ability to perform a functional grasping task. The ball-grasp device was equipped with a contact sensor and a task-success sensor that were activated when the monkeys initially touched the ball and dropped it into the tube, respectively.

[1]Department of Physiology, Feinberg School of Medicine, Northwestern University, 303 East Chicago Avenue, Chicago, Illinois 60611, USA. [2]Department of Bioengineering, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA. [3]Department of Biomedical Engineering, Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208, USA. [4]Department of Physical Medicine and Rehabilitation, Feinberg School of Medicine, Northwestern University, 345 East Superior Avenue, Chicago, Illinois 60611, USA.

were well-modulated during at least some portion of the task. Offline, typically 50–75% of the neuronal signals could be discriminated as single neurons, on the basis of the consistency of their waveform shape and inter-spike interval histogram distribution. However, under real-time conditions, only about one-third of the inputs were well-discriminated single units; the remainder were signals that included action potentials from more than one neuron. Figure 2b shows the discharge of these neurons averaged over 229 trials, aligned to the time of contact with the ball. The varied phasing of the different neurons is evident.

We recorded from flexor and extensor muscles of the hand and forearm simultaneously with the neural recordings (Fig. 2c, d). There was considerable variation both in the magnitude and duration of electrical activity that occurred from trial to trial (Fig. 2c), and in the average timing and patterns of activation of the different muscles (Fig. 2d).

We were able to predict electromyographic (EMG) activity with very high accuracy, typically from approximately 100 neural signals (Fig. 2c, d; red traces), using Wiener cascade decoders. These decoders consisted of multiple-input, linear-impulse response functions between the neural inputs and each muscle, followed by a static non-linearity. Each impulse response was composed of ten lags spanning 500 ms. At the beginning of each week, we collected 20 min of data under normal conditions, and we used this to compute the coefficients for the decoder that were then used for the remainder of the week. Accuracy was represented by R^2, calculated using a multi-fold cross-validation procedure described in the Supplementary Information.

Using these real-time predictions of muscle activity, we stimulated up to five electrodes in three different muscles (flexor carpi radialis (FCR), medial and lateral sites in the flexor digitorum superficialis (FDS) and flexor digitorum profundus (FDP)). By these means, we have restored in two monkeys the ability to pick up and move objects despite complete paralysis of the flexor muscles in the forearm and hand. We began each FES experimental session by collecting data under normal conditions to establish baseline performance. Following these baseline recordings, we injected lidocaine through nerve cuffs implanted proximal to the elbow that blocked the median and ulnar nerves. After 15–20 min the nerve block was complete, as determined by the loss of flexor muscle EMG activity (see Supplementary Information and Supplementary Fig. 1), and the onset of profound motor deficits. We made periodic tests of nerve-block effectiveness throughout each session (Supplementary Fig. 2), and we used a standardized stimulus train to evaluate the level of fatigue induced by the stimulation (Supplementary Fig. 2).

A series of four trials is shown in Fig. 3a, b, showing typical neural discharge, predicted EMG and stimulus commands, as well as markers of the monkey's performance. Although the common digit flexors (FDS and FDP) are normally activated nearly synchronously, FDS activation tended to be more sustained, whereas FDP was more phasic. The pulse widths of the stimulus trains used to activate a given muscle

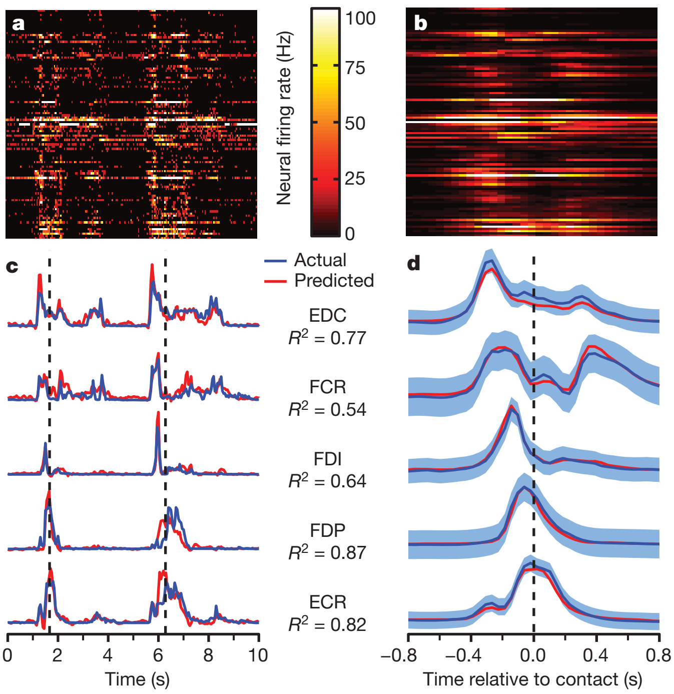


Figure 2 | Grasp-related raw data collected during normal conditions. **a**, Firing rates of 104 neuronal signals recorded during series of two grasps. **b**, Ensemble average of 229 trials aligned to the time of ball contact. **c**, Actual and predicted EMG during the same period as (**a**), with the muscles ordered by the relative times of their onset, including extensor digitorum communis (EDC), flexor carpi radialis (FCR), first dorsal interosseous (FDI), flexor digitorum profundus (FDP), and extensor carpi radialis (ECR). Predicted EMG was computed using multiple-input linear-impulse response decoders built from data collected earlier in the session. Vertical dashed lines mark the times of ball contact. R^2 values indicate prediction accuracy for the 20-min data file. **d**, Ensemble averages of EMG activity, aligned to the time of initial contact. Blue shaded regions, ±1 s.d. around the mean.

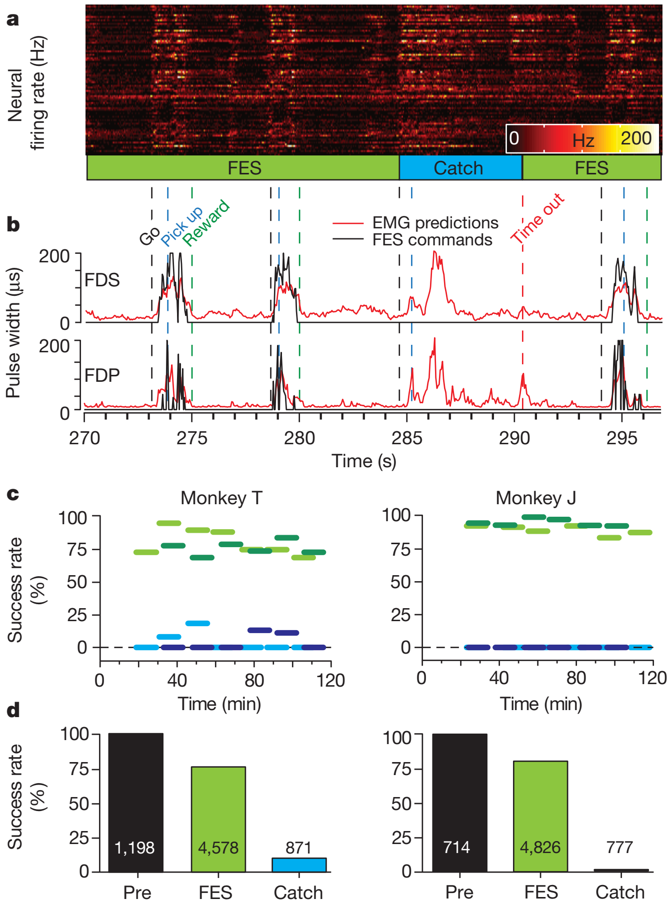


Figure 3 | Grasp performance during four consecutive brain-controlled FES trials. **a**, Neural data. **b**, Predicted EMG signals (red traces) transformed into stimulus commands (black traces). Vertical dashed lines: go tone (Go), time of initial ball contact (Pick up) and successful task completion (Reward). **c**, Horizontal lines show average success rates for sequential 10-min blocks during two separate experimental sessions (indicated by light and dark horizontal lines, respectively). Each session included both FES trials (green lines) and catch trials without stimulation (blue lines). The neuroprosthesis markedly improved the monkeys' ability to grasp the ball despite paralysis. **d**, Average success rates for pre-block (Pre), FES and catch (Catch) trials across all sessions (100%, 76% and 10% for Monkey T; 99%, 80% and 1% for Monkey J). The total number of trials (successful and failed) is displayed on the bars for each condition.

were determined from the predicted EMG for that muscle using a mapping procedure described in the Supplementary Information and Supplementary Fig. 3. The distribution of these pulse widths throughout the full range from 0 to 200 μs suggests that the monkey was able to grade the strength of contraction continuously (Supplementary Fig. 4). During the FES trials, the monkey grasped and moved the ball reliably. The movements did not differ sufficiently from normal to be obvious to casual observation (see Supplementary Movies 1 and 2 for representative examples from both monkeys). On occasional 'catch' trials, we turned off the neuroprosthesis at the beginning of the trial, to test the ability of the monkey without FES. In the example of a catch trial illustrated here (note the flat stimulus trace in Fig. 3b), the monkey was unable to grasp the ball despite the considerable effort apparent in the neural discharge and predicted EMG.

After the onset of paralysis, each experimental session consisted of a series of 10-min sets of trials like those in Fig. 3, in which the monkey attempted to complete the grasp task either with or without FES assistance. Two complete sessions for both monkeys are summarized by the horizontal light and dark lines in Fig. 3c. The success rate in these sessions using the neuroprosthesis was approximately 80% and 90% for the two monkeys, respectively (green lines). In contrast, the average catch-trial success rate was 5% for monkey T and 0% for monkey J (blue lines). The average number of trials per session varied substantially across sessions, with a mean of 272 ± 84 for monkey T, and 208 ± 112 for monkey J. Although we tried different types of balls, we did not systematically examine the effects of size, weight or texture on the monkeys' performance. It is likely that the FES success rate would have been lower if balls that were substantially heavier or more slippery had been used. We chose to use balls that in size and weight mimicked objects grasped in routine human tasks (for example, eating an apple).

Figure 3d summarizes both monkeys' overall success rate across all sessions, both with the FES neuroprosthesis and during catch trials. Both monkeys achieved a success rate of approximately 80% using the neuroprosthesis, a level that was highly significantly different ($P < 0.0001$) from that of the catch-trial condition. In addition to resulting in a greatly improved success rate, the FES neuroprosthesis also significantly increased the speed at which the monkeys completed successful trials (not shown; $P < 0.0001$ for both monkeys, two-tailed Mann–Whitney test).

To test force control more systematically we conducted a second set of experiments with monkey J, who was trained to control the vertical displacement of a cursor that moved in proportion to palmar grasp force. Using the neuroprosthesis, the monkey was able to squeeze a pneumatic tube, and to track up to three different targets ranging from 15 to 80% of his normal maximum voluntary contraction (MVC), each target having a width of approximately 20% of MVC. To be successful, the monkey needed to maintain the target force for 0.5 s. Figure 4 shows a short sequence of data during this target tracking task. One of these four trials was a catch trial. The monkey was unable to generate any force during the catch trial despite two attempts that are evident in the predicted EMG signals.

We quantified this performance by measuring the mean force and stimulation pulse width during the target-hold periods of the initial and final 10 min of the session. Despite considerable FES-induced fatigue, the monkey remained able to achieve the required force throughout the session by voluntarily increasing the mean stimulus pulse width (see Supplementary Fig. 5). The increased pulse width reflects an increase in cortical activity and resultant EMG predictions. The monkey seemed to overcome the fatigue in a manner similar to that of normal conditions, increasing its effort to regulate force accurately.

The monkey's ability to control both a well-regulated palmar grasp as well as to execute the unconstrained natural grasp is powerful evidence of the impact that this FES neuroprosthesis could have in eventual clinical application. Our neuroprosthesis makes use of patterns of activity in M1 that reflect the patterns that occur naturally during grasp. By matching patterns of neuronal activity to those muscles with

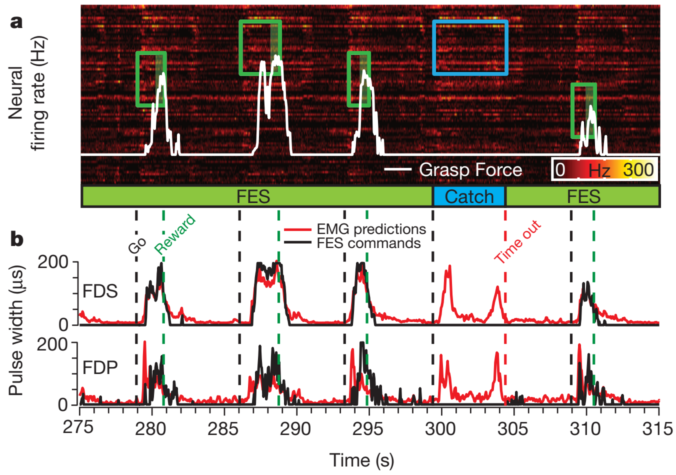


Figure 4 | FES used to produce controlled palmar grasp force during the palmar grasp task. a, Colour-coded neural discharge recorded during a series of five trials. The rectangles indicate times of target appearance and disappearance, as well as their upper and lower force bounds. The white trace is the force generated by the monkey, resulting from stimulation of FDS and FDP (Stim commands). **b**, There were four successful trials with FES (green targets) and one unsuccessful catch trial (blue target). During the catch trial, the monkey made two unsuccessful attempts to squeeze the tube, as seen in the EMG predictions (also evident in the neural discharge (**a**)).

which they are normally most closely correlated, we hope to maintain the natural coupling between cortical activity and motor output.

It is important to note that this process in no way limits the ability of the brain to adapt further, to compensate for inaccuracies in the decoded signals or the stimulus-evoked contractions. Even with adaptation, it is difficult to imagine how a small number of individually conditioned, randomly selected neurons could yield an adequate level of control without the type of pre-programming that is necessary with existing FES systems. Indeed, there is no evidence that it is possible to learn to associate the simultaneous activity of two, three or more neurons with independent patterns of muscle activity. Even if possible, the cognitive load associated with this effort would presumably be rather high, whereas the reliability of a neuroprosthesis relying on a small number of conditioned neurons would be quite low.

Our model of paralysis avoided many of the complications of actual spinal cord injury, including muscle denervation and spasticity[19,20]. Furthermore, it was limited to the forearm and digit flexors. Patients with C5 and C6 spinal cord injury retain voluntary control of proximal arm muscles while losing full control of the more distal limb. Many patients retain or regain some level of voluntary wrist extension[21]. As we did not paralyse the monkeys' extensor muscles in this experiment, it is important to recognize the good coordination between the remaining natural muscle control and that achieved through the neuroprosthesis. We routinely obtained extensor EMG predictions that were in fact slightly more accurate than those of the flexors. In future experiments, we intend to expand our control to these muscles.

This technology may offer even greater advantages to patients with more severe injuries, who have a greater need for replaced function but possess even fewer available sources of control[22]. In addition to the ability to predict the reach-related activity of distal limb muscles considered in this study, we have previously showed the same ability in relation to proximal limb muscles, suggesting the possibility of extending this control to these muscles[13]. As well as offering patients greater independence, FES is also established as an effective means for exercising the muscles of paralysed patients, bringing a range of health benefits: stronger muscles and bones, improved metabolism, cardiorespiratory health and reduced propensity to pressure sores[23,24]. It may be that drawing on a conscious process to restore natural movement will bring the additional benefit of improved psychological health[25].

METHODS SUMMARY

Experimental subjects and task. Two monkeys were trained to perform a ball-grasp task (Fig. 1) and one of the monkeys was also trained to perform a controlled-force palmar grasp task. The monkeys were allowed 5 s to grasp one of several balls (ranging in size from 25–40 mm in diameter and 55–130 g) and place it into the top of a dispenser tube. The palmar grasp task required the monkey to squeeze a pneumatic tube that controlled movement of a cursor. Force targets were chosen from a set of two or three non-overlapping levels. All procedures were approved by the Institutional Animal Care and Use Committee of Northwestern University, Illinois, USA.

EMG prediction. Inputs consisted of roughly 100 single and multi-unit signals from a 100-electrode array (Blackrock Microsystems) implanted within the hand area of M1. Decoders consisted of multiple-input impulse response functions between the neural inputs and each muscle, subsequently transformed by a second-order static nonlinearity to reduce the baseline noise in the predictions and to increase the gain near the EMG peaks[13,16]. We computed decoders at the beginning of each week, which were used in daily sessions for the remainder of the week. We conducted 20 sessions with 7 decoders across 7 weeks for monkey T and 27 sessions with 6 decoders across 11 weeks with monkey J.

Stimulation. All muscles were stimulated at a single, fixed rate of either 25 or 30 Hz to achieve nearly fused contractions. The EMG predictions were transformed into stimulus pulse widths by mapping the predicted EMG noise floor to the stimulus force threshold, and the maximum predicted EMG to the maximum pulse width (200 μs; see Supplementary Fig. 3). The current, typically 2–8 mA, was chosen independently for each electrode, to yield forces of approximately 50% of the maximal evocable force at 200-μs pulse width.

Received 23 December 2011; accepted 22 February 2012.
Published online 18 April 2012.

1. Keith, M. W. *et al.* Implantable functional neuromuscular stimulation in the tetraplegic hand. *J. Hand Surg. Am.* **14,** 524–530 (1989).
2. Peckham, P. H. *et al.* Efficacy of an implanted neuroprosthesis for restoring hand grasp in tetraplegia: a multicenter study. *Arch. Phys. Med. Rehabil.* **82,** 1380–1388 (2001).
3. Pohlmeyer, E. A., Jordon, L. R., Kim, P. & Miller, L. E. A fully implanted drug delivery system for peripheral nerve blocks in behaving animals. *J. Neurosci. Methods* **182,** 165–172 (2009).
4. International. Campaign for Cures of Spinal Cord Injury Paralysis. http://www.campaignforcure.org (2011).
5. Anderson, K. D. Targeting recovery: priorities of the spinal cord-injured population. *J. Neurotrauma* **21,** 1371–1383 (2004).
6. Popovic, M. R., Popovic, D. B. & Keller, T. Neuroprostheses for grasping. *Neurol. Res.* **24,** 443–452 (2002).
7. Kilgore, K. L. *et al.* An implanted upper-extremity neuroprosthesis. Follow-up of five patients. *J. Bone Joint Surg. Am.* **79,** 533–541 (1997).
8. Evarts, E. V. Relation of pyramidal tract activity to force exerted during voluntary movement. *J. Neurophysiol.* **31,** 14–27 (1968).
9. Humphrey, D. R., Schmidt, E. M. & Thompson, W. D. Predicting measures of motor performance from multiple cortical spike trains. *Science* **170,** 758–761 (1970).
10. Carmena, J. M. *et al.* Learning to control a brain–machine interface for reaching and grasping by primates. *PLoS Biol.* **1,** e42 (2003).
11. Gupta, R. & Ashe, J. Offline decoding of end-point forces using neural ensembles: application to a brain–machine interface. *Neural Systems and Rehabilitation Engineering. IEEE Trans. Neural Syst. Rehabil. Eng.* **17,** 254–262 (2009).
12. Fagg, A. H., Ojakangas, G. W., Miller, L. E. & Hatsopoulos, N. G. Kinetic trajectory decoding using motor cortical ensembles. *IEEE Trans. Neural Syst. Rehabil. Eng.* **17,** 487–496 (2009).
13. Pohlmeyer, E. A., Solla, S. A., Perreault, E. J. & Miller, L. E. Prediction of upper limb muscle activity from motor cortical discharge during reaching. *J. Neural Eng.* **4,** 369–379 (2007).
14. Cherian, A., Krucoff, M. O. & Miller, L. E. Motor cortical prediction of EMG: evidence that a kinetic brain-machine interface may be robust across altered movement dynamics. *J. Neurophysiol.* **106,** 564–575 (2011).
15. Pohlmeyer, E. A. *et al.* Real-time control of the hand by intracortically controlled functional neuromuscular stimulation. *IEEE 10th Int. Conf. Rehab. Robotics* 454–458 (2007).
16. Pohlmeyer, E. A. *et al.* Toward the restoration of hand use to a paralyzed monkey: Brain-controlled functional electrical stimulation of forearm muscles. *PLoS ONE* **4,** e5924 (2009).
17. Oby, E. R. *et al.* in *Statistical Signal Processing for Neuroscience and Neurotechnology* (ed. O'Weiss, K.G.) 369–406 (Academic Press, 2010).
18. Moritz, C. T., Perlmutter, S. I. & Fetz, E. E. Direct control of paralysed muscles by cortical neurons. *Nature* **456,** 639–642 (2008).
19. Adams, M. M. & Hicks, A. L. Spasticity after spinal cord injury. *Spinal Cord* **43,** 577–586 (2005).
20. Kern, H. *et al.* Denervated muscles in humans: limitations and problems of currently used functional electrical stimulation training protocols. *Artif. Organs* **26,** 216–218 (2002).
21. Waters, R. L., Adkins, R. H., Yakura, J. S. & Sie, I. Motor and sensory recovery following complete tetraplegia. *Arch. Phys. Med. Rehabil.* **74,** 242–247 (1993).
22. Bryden, A. M. *et al.* An implanted neuroprosthesis for high tetraplegia. *Top. Spinal Cord Inj. Rehabil.* **10,** 38–52 (2005).
23. Nightingale, E. J., Raymond, J., Middleton, J. W., Crosbie, J. & Davis, G. M. Benefits of FES gait in a spinal cord injured population. *Spinal Cord* **45,** 646–657 (2007).
24. Agarwal, S. *et al.* Long-term user perceptions of an implanted neuroprosthesis for exercise, standing, and transfers after spinal cord injury. *J. Rehabil. Res. Dev.* **40,** 241–252 (2003).
25. Fitzwater, R. A personal user's view of functional electrical stimulation cycling. *Artif. Organs* **26,** 284–286 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported in part by grant NS053603 from the National Institute of Neurological Disorders and Stroke to L.E.M. and a post-doctoral fellowship from the Fonds de la Recherche en Santé du Québec to C.E., with further support from the Chicago Community Trust through the Searle Program for Neurological Restoration at the Rehabilitation Institute of Chicago. We also acknowledge the technical assistance of D. Tyler and K. Kilgore as well as the surgical assistance of J. Ko and S. Paisley Agnew.

Author Contributions L.E.M. conceived, designed and supervised the basic experiments. C.E. and E.R.O. performed the experiments. M.J.B. carried out software development. C.E. analysed the data and prepared figures. L.E.M. and C.E. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to L.E.M. (lm@northwestern.edu).

LETTER

doi:10.1038/nature11091

KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant

Christelle Golzio[1], Jason Willer[1], Michael E. Talkowski[2,3], Edwin C. Oh[1], Yu Taniguchi[4], Sébastien Jacquemont[5], Alexandre Reymond[6], Mei Sun[2], Akira Sawa[4], James F. Gusella[2,3], Atsushi Kamiya[4], Jacques S. Beckmann[5,7] & Nicholas Katsanis[1,8]

Copy number variants (CNVs) are major contributors to genetic disorders[1]. We have dissected a region of the 16p11.2 chromosome—which encompasses 29 genes—that confers susceptibility to neurocognitive defects when deleted or duplicated[2,3]. Overexpression of each human transcript in zebrafish embryos identified *KCTD13* as the sole message capable of inducing the microcephaly phenotype associated with the 16p11.2 duplication[2–5], whereas suppression of the same locus yielded the macrocephalic phenotype associated with the 16p11.2 deletion[5,6], capturing the mirror phenotypes of humans. Analyses of zebrafish and mouse embryos suggest that microcephaly is caused by decreased proliferation of neuronal progenitors with concomitant increase in apoptosis in the developing brain, whereas macrocephaly arises by increased proliferation and no changes in apoptosis. A role for *KCTD13* dosage changes is consistent with autism in both a recently reported family with a reduced 16p11.2 deletion and a subject reported here with a complex 16p11.2 rearrangement involving *de novo* structural alteration of *KCTD13*. Our data suggest that *KCTD13* is a major driver for the neurodevelopmental phenotypes associated with the 16p11.2 CNV, reinforce the idea that one or a small number of transcripts within a CNV can underpin clinical phenotypes, and offer an efficient route to identifying dosage-sensitive loci.

Copy number changes have emerged as a notable source of genetic variation contributing to the human genetic disease risk[1]. In addition to genomic disorders such as Charcot–Marie–Tooth disease, DiGeorge syndrome and others[7,8], in which large deletions and duplications represent penetrant alleles for discrete syndromic phenotypes, recent advances have highlighted the contribution of such genomic events in a broad range of both common and rare traits (see the DECIPHER consortium website, http://decipher.sanger.co.uk). Systematic surveys of neurodevelopmental disorders have uncovered a particularly high incidence of both inherited and *de novo* CNVs that can confer either causality or susceptibility[9–12]. For example, deletions in 1q21.1 and 15q13.3 have been associated with schizophrenia, whereas duplications in 15q11–15q13 and 7q22–7q31 have been associated with autism spectrum disorder (ASD; see review[1]).

A 600-kb deletion on 16p11.2, encompassing 29 annotated genes, has been significantly and reproducibly associated with a range of neurocognitive defects, including epilepsy, autism and ASD[2], whereas the reciprocal duplication has been associated with autism and schizophrenia[3]. In addition, extended phenotypic analyses of patients with such genomic lesions have revealed strong mirroring comorbidities: the common 16p11.2 deletion is associated with paediatric neurodevelopmental disorders, including autism, diabetes-independent obesity[5] and macrocephaly[6], whereas the reciprocal duplication is associated with both autism and schizophrenia, as well as anorexia and microcephaly[2–5,13]. Moreover, a recent *post hoc* analysis of ASD and schizophrenia loci has revealed that such comorbidities might be causally linked to each other, with macrocephaly shown to be associated with ASD, and microcephaly with schizophrenia[31].

A pervasive challenge in the interpretation of CNV discovery is the transition from the detection of a genomic lesion that can often span large regions encompassing many genes to the identification of the critical loci whose dosage sensitivity drives the phenotype. For some disorders, this has been achieved through the discovery of highly penetrant point mutations at a single locus; for example, mutations in *PMP22* are sufficient to cause CMT[14], whereas mutations in *RAI1* cause Smith–Magenis syndrome[15]. In other disorders, gene-specific genomic alterations such as chromosomal translocations, inversions or small coding deletions can narrow the critical region to a single gene (for example, *MBD5* in 2q23.1 microdeletion syndrome[16]). Alternatively, systematic functional dissection through mouse mutagenesis has yielded strong candidates; ablation of *Tbx1* recapitulates the cardiac phenotypes of VCSF[17]; similarly, knockout of *Shank3* captures most of the phenotypes seen in the terminal 22q deletion that causes Phelan–McDermid syndrome[18]. However, these approaches are considerably more challenging for common phenotypes and genetically heterogeneous disorders: systematic engineering of the mouse genome for each gene in a CNV can be impractical, and rare mutations involved in complex traits are likely to exhibit both reduced penetrance and variable expressivity.

Manipulation of zebrafish embryos is an attractive alternative method to discover human dosage-sensitive genes, particularly when the CNV under investigation has mirrored anatomical phenotypes that are detectable during early development and that can therefore be assayed using a combination of gene suppression and overexpression experiments[19]. Given the association between the 16p11.2 CNV and changes in head size, we proposed that (1), systematic overexpression of each of the 29 genes in the common duplication might yield a defined, reproducible set of transcripts, some of which might cause microcephaly; and (2) reciprocal suppression of these genes should yield the macrocephalic phenotype seen in the 16p11.2 deletion. To test these possibilities, we first queried the zebrafish genome by reciprocal BLAST (basic local alignment search tool) for each of the 29 target genes (Fig. 1a) and identified 24 orthologues (Supplementary Table 1), with five genes, *SPN, QPRT*, *C16orf54*, *TMEM219* and *C16orf92* found only in placental mammals. In a manner akin to the classic *Drosophila* misexpression experiments, we generated capped messenger RNA for all 29 human genes and injected zebrafish embryos at the two-cell stage with equimolar pairwise 'cocktail' combinations at two dosages of 25 pg and 50 pg. These commonly used ranges[20] were selected because they represent $>$0.25–0.5% of total polyA^{+} mRNA in a zebrafish embryo[21] and are therefore likely to achieve substantial overexpression above the baseline of any single transcript.

One gene, *TAOK2*, required a reduction in mRNA dosage to 10 pg because of toxicity. For the remaining 28 genes, we observed no

[1]Center for Human Disease Modeling and Department of Cell biology, Duke University, Durham, North Carolina 27710, USA. [2]Molecular Neurogenetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [3]Departments of Neurology and Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland 21287, USA. [5]Service of Medical Genetics, Centre Hospitalier Universitaire Vaudois, 1011 Lausanne, Switzerland. [6]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland. [7]Department of Medical Genetics, University of Lausanne, 1005 Lausanne, Switzerland. [8]Department of Pediatrics, Duke University, Durham, North Carolina 27710, USA.

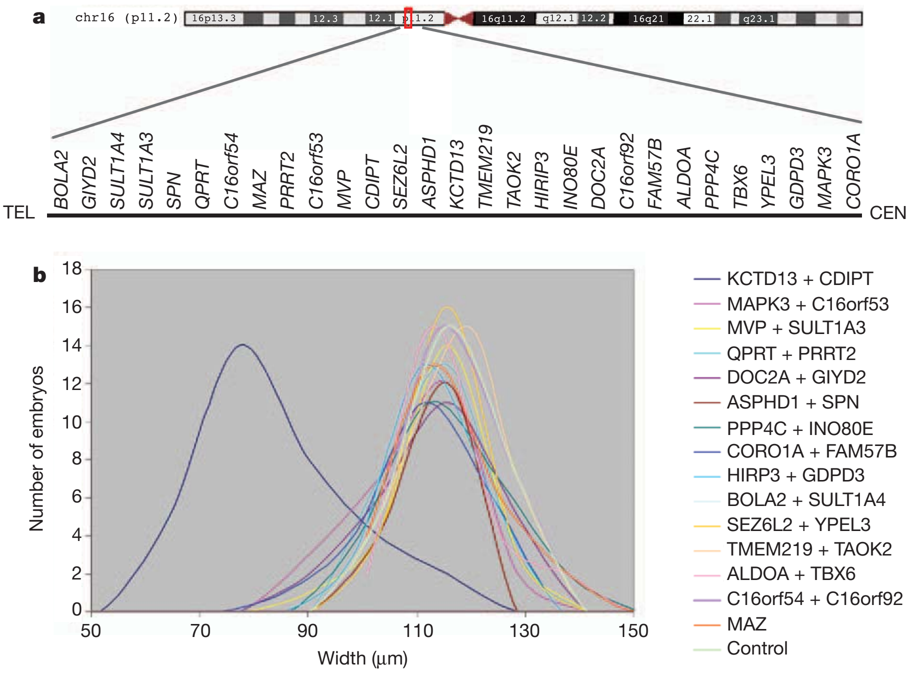


Figure 1 | Systematic analysis of 16p11.2 deletion or duplication genes *in vivo* induces defects in head size. **a**, Schematic of chromosome 16, with detail of the 16p11.2 CNV, showing gene content (not to scale) above the black line. **b**, Plot of head size measurements (in μm) of human mRNA overexpression combinations measured across approximately 50 embryos per injection cocktail. In all but one injection, controls and human overexpressed genes resulted in indistinguishable median head size, with minimal variance. In contrast, embryos injected with the *KCTD13* and *CDIPT* mRNA cocktail show consistent and significant reduction of head size.

lethality or gross morphological defects at either 25 pg or 50 pg: random PCR with reverse transcription (RT–PCR) testing of nine injection cocktails, including *KCTD13*, showed persistence of the corresponding human mRNAs up to approximately 4.25 days post fertilization (d.p.f.) (Supplementary Fig. 1e). We therefore developed a surrogate measurement for head size at 4.25–4.5 d.p.f. using objective measurements, with the distance across the convex tips of the eye cups recorded in 50 embryos per injection, scored by an investigator who was uninformed of the composition of the injection cocktail (Fig. 1b). Only a single overexpression cocktail containing *KCTD13* and *CDIPT* gave significant changes in head size (two-tailed t-test, $P < 0.000001$). Subsequent single-mRNA injections for the two genes indicated that the phenotype was driven exclusively by the overexpression of *KCTD13*; injection of *KCTD13* at progressively increasing mRNA amounts yielded an increasing percentage of microcephalic embryos (Fig. 2a, b and Supplementary Fig. 1a). In contrast, the head size of embryos injected with *CDIPT* was indistinguishable from those injected with sham control (data not shown).

To validate the specificity of this phenotype and to investigate whether it was possible to also simulate the macrocephalic phenotype seen usually in 16p11.2del patients[4,13], we designed a splice-blocking morpholino against the donor site of exon 3 of the sole *kctd13* zebrafish orthologue. Injection of 10 ng of morpholino followed by RT–PCR testing showed ~70% reduction of *kctd13* message at 4.5 d.p.f. (Supplementary Fig. 1c). Notably, this injection also yielded a significant increase in mean head size ($P < 0.00001$; Fig. 2a, b and Supplementary Fig. 1b). This phenotype is specific to *kctd13*; a scrambled morpholino induced no phenotypes, whereas injection of 10 ng of morpholino and 50 pg of *KCTD13* mRNA rescued both the microcephalic and macrocephalic phenotypes (Supplementary Fig. 1d). Importantly, measurement of the somitic trunk length of scored embryos showed no differences in length (or morphology; Supplementary Fig. 2), indicating that the head size differences are unlikely to be driven by gross developmental delay. There were also no defects in other structures, including the heart and the swim bladder.

To investigate the mechanism of the head size defects, we examined the developing brain of both macro- and microcephalic embryos. *In situ* hybridization with an antisense *kctd13* probe showed that this transcript is expressed strongly in the developing brain. At 24 hours post fertilization (h.p.f.), *kctd13* is strongly expressed in the anterior forebrain (Supplementary Fig. 3a, f), the midbrain and the hindbrain. In later stages, *kctd13* is expressed predominantly in the telencephalon, the diencephalon and the retina (Supplementary Fig. 3b–e). Staining both phenotype classes with terminal deoxynucleotidyl TdT-mediated dUTP nick end labelling (TUNEL) showed a significant increase in apoptosis exclusively in the microcephalic embryos. At the same time, phospho-histone H3 antibody staining revealed an increase in proliferating cells in the brain of *kctd13* morphants and a reciprocal decrease of proliferating cells in overexpressant embryos (Fig. 2 c, d). Detailed analysis of transverse sections from embryos injected with either morpholino or *KCTD13* mRNA confirmed that the observed phenotypes are probably driven by changes in the number of cells in the developing brain; the architecture and cell content of the Meckel's and palatoquadrate pharyngeal cartilages were normal, as was the overall cell content and architecture of the retina (Fig. 3a–f). Cell nuclei were counted at 4.5 d.p.f. (at the stage when the anatomical measurements were made) and this showed that there were significant reciprocal changes in the total number of cells in each of the telencephalon, diencephalon and mesencephalon (Fig. 3k). Furthermore, counting of HuC/D-positive cells (a marker for post-mitotic neurons; cells were positive for HuC and HuD (also known as ELAVL3 and ELAVL4)) in the telencephalon recapitulated the differences seen in total cell count but showed no difference in cell circumference (Fig. 3l and Supplementary Fig. 4), indicating that the changes in overall cell numbers and ultimate changes in head size are largely driven by changes in the numbers of mature neurons. These data, collected at 4.5 d.p.f., predict that the onset of the neuroanatomical defects would precede the manifestation of anatomical micro- or macrocephalic phenotypes. To test this possibility, we stained all three classes of embryos with HuC/D and acetylated tubulin at 2 d.p.f., by which time the head size of *kctd13* morpholinos and overexpressants is indistinguishable from uninjected controls. We observed stark differences in the density and distribution of neurons, particularly in the forebrain, with concomitant loss of organization and bilateral symmetry (Fig. 3m), as well as aberrant distribution of axonal tracks (Fig. 3n).

Although the zebrafish brain bears many similarities with the mammalian brain in terms of developmental programming, we examined whether our findings might be relevant to cortical development in a mammalian system. Data mining of the Gene Expression Nervous System Atlas (GENSAT; http://www.gensat.org) and the Allen brain atlas (http://www.brain-map.org) revealed that human and mouse *KCTD13* is expressed throughout development in neurons residing in the cortex, striatum, olfactory tubercle and hippocampus. We

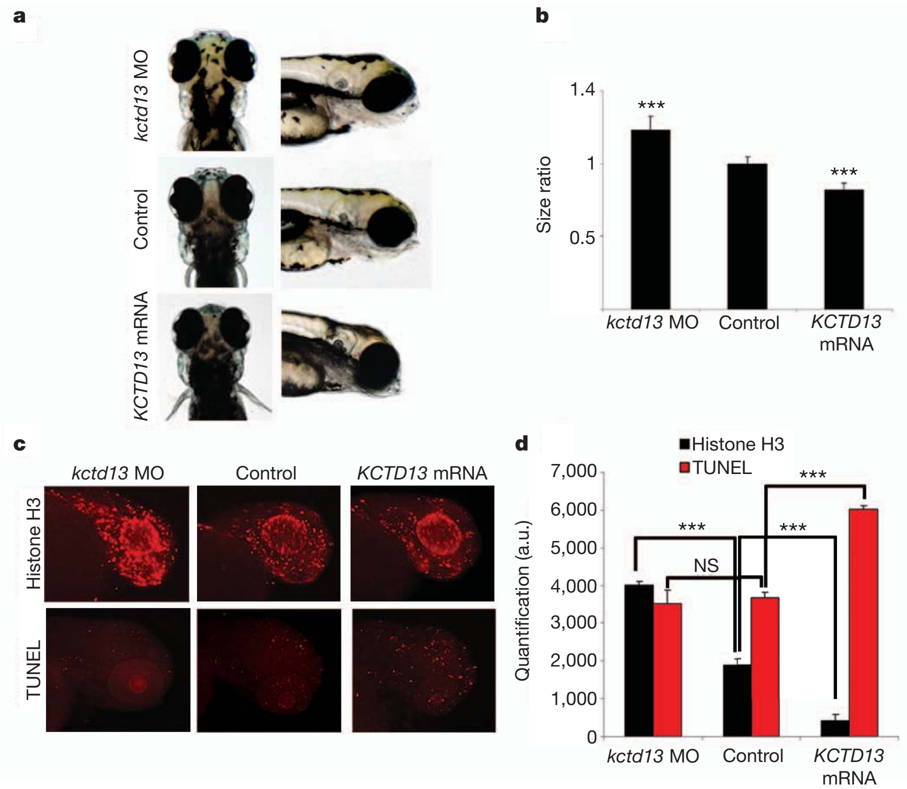


Figure 2 | KCTD13 dosage changes lead to head size, proliferation and apoptosis defects. **a**, Dorsal (left) and lateral (right) views of representative embryos injected with (from top to bottom) *kctd13* morpholino (MO), control or *KCTD13* mRNA. **b**, Graph of the ratio between head size of control and injected embryos at 4.5 d.p.f. ($n = 45$ per injection). **c**, Phospho-histone H3 and TUNEL staining for proliferating or apoptotic cells in the zebrafish brain at 2 d.p.f. and 3 d.p.f., respectively. Representative examples of MO-, control- and mRNA-injected embryos are shown. **d**, Graph of phospho-histone H3 and TUNEL quantifications from 20 MO-, control- and mRNA-injected embryos. Data from three independent experiments are represented as mean ± s.d. ***$P < 0.00001$; two-tailed t-test comparisons between control and either MO- or mRNA-injected embryos. a.u., arbitrary units; NS, not significant.

therefore designed short hairpin RNAs (shRNAs) against murine *Kctd13* and tested their efficiency in a cultured mouse neuroblastoma cell line (Neuro-2a). One shRNA, which, similar to the *kctd13* morpholino, downregulated the expression of endogenous message by ~70% (Supplementary Fig. 5), was then co-transfected with a green fluorescent protein (GFP)-expressing plasmid into Neuro-2a cells. Two days after transfection, cells were pulsed with 5-bromodeoxyuridine (BrdU) and analysed for the effects of *Kctd13* knockdown on cellular proliferation. Similar to the proliferation data from zebrafish morphants, depletion of *Kctd13* resulted in a 34% increase ($P < 0.01$) in BrdU/GFP-labelled cells (Fig. 4a). Next, we injected the *Kctd13* shRNA and GFP-expressing plasmid into the ventricular space of wild-type C57BL/6 embryos at embryonic day 13.5 (E13.5) and injected BrdU into pregnant dams 2 h before collection of electroporated embryos at E15.5. Knockdown of *Kctd13* resulted in a twofold increase in BrdU/GFP labelling within the ventricular zone ($P < 0.001$; Fig. 4b), suggesting that *Kctd13* is required to maintain the proliferative status of cortical progenitors *in vivo*.

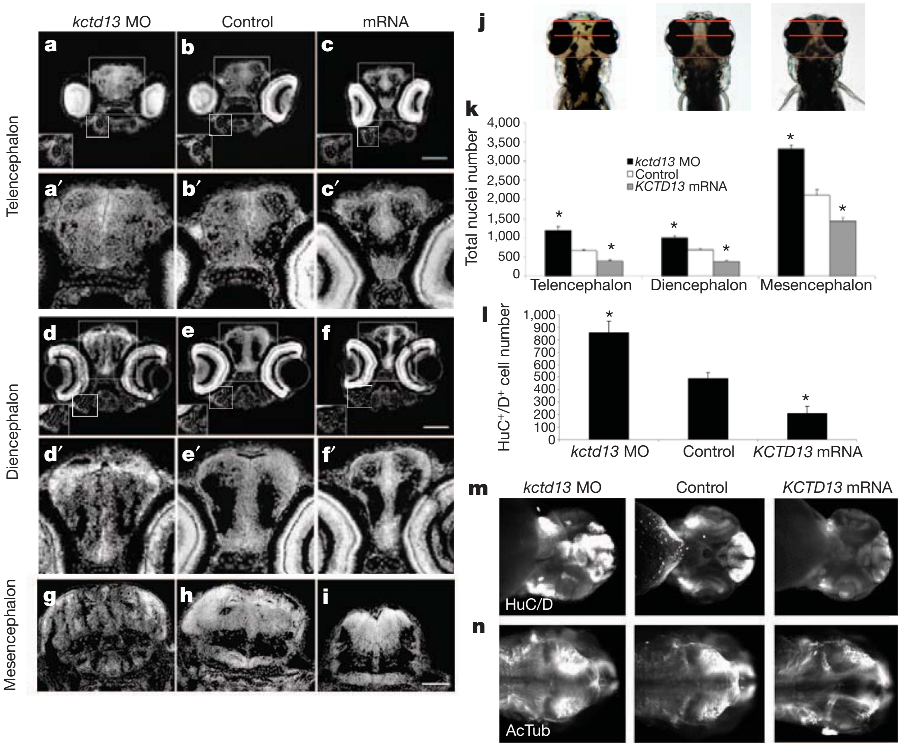


Figure 3 | KCTD13 dosage changes lead to neuroanatomical defects. **a–c**, DAPI (4′,6-diamidino-2-phenylindole) staining on transverse sections of the telencephalon of embryos injected with *kctd13* MO, control or KCTD13 mRNA at 4.5 d.p.f. The Meckel's pharyngeal cartilage is shown in the insets. Higher magnifications are shown in **a′–c′**. **d–f**, Same as **a–c** but for the diencephalon. The palatoquadrate pharyngeal cartilage is shown in the insets. Higher magnifications are shown in **d′–f′**. **g–i**, Same as **a–c** but for the mesencephalon. **j**, The planes of section are illustrated with red lines on dorsal views of kctd13 MO-, control- and *KCTD13* mRNA-injected embryos (left to right). **k**, Bar graph of the total number of nuclei for the three classes of embryos in the telencephalon, diencephalon and mesencephalon at 4.5 d.p.f. (3 adjacent sections, $n = 4$). **l**, Bar graph of the number of HuC/D-positive (HuC^+/D^+) cells in the telencephalon for *kctd13* MO-, control- and *KCTD13* mRNA-injected embryos at 4.5 d.p.f. (3 adjacent sections, $n = 4$). **m**, **n**, Ventral (**m**) and dorsal views (**n**) of *kctd13* MO-, control- and *KCTD13* mRNA-injected embryos at 2 d.p.f stained with either anti-acetylated tubulin (AcTub) or HuC/D. Data are represented as mean ± s.d. *$P < 0.01$; two-tailed t-test comparisons between control and either MO- and mRNA-injected embryos. Scale bars, 100 µm.

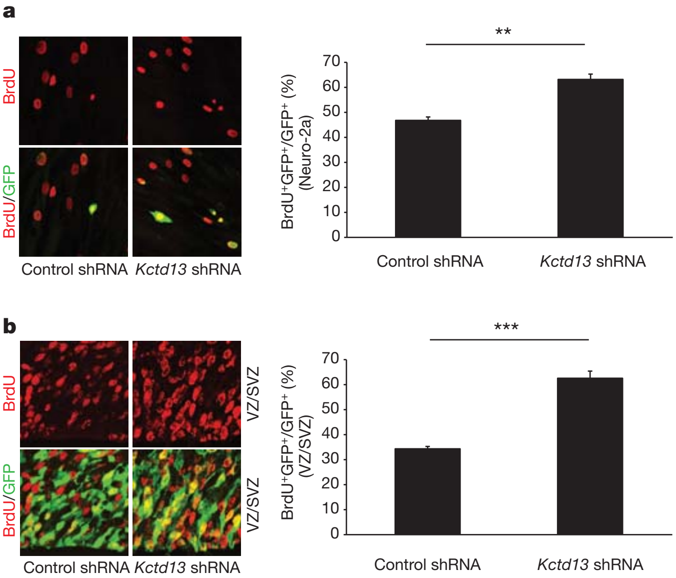


Figure 4 | *Kctd13* regulates mammalian cell proliferation *in vitro* and *in vivo*. **a**, Knockdown of *Kctd13* in Neuro-2a cells results in an increase in the number of BrdU- and GFP-positive ($BrdU^+GFP^+$) cells relative to control cells (GFP^+). Error bars represent the standard error from two independent experiments. **b**, Analysis of E15.5 mouse cortices injected with either *Kctd13* or control shRNA reveal a similar increase in $BrdU^+GFP^+$ cells in knockdown tissue ($n = 4$). Error bars represent the standard error from two independent experiments. SVZ, subventricular zone; VZ, ventricular zone. $**P < 0.01$; $***P < 0.001$.

The combination of our zebrafish and mouse data suggests that *KCTD13* is a major driver of the head size phenotypes associated with the 16p11.2 CNV through the regulation of early neurogenesis. Our data do not exclude the possibility that other loci also have an independent contribution to the 16p11.2 deletion or duplication (16p11.2del/dup) anatomical phenotypes, and cannot directly answer the question of whether all of the observed pathology in 16p11.2del/dup patients is driven by dosage changes in *KCTD13*. However, we were able to ask whether dosage changes of other loci inside the CNV might also be relevant to the head size phenotypes established by *KCTD13*. Specifically, we performed pairwise overexpressions of *KCTD13* with each of the other 28 transcripts in the region and asked whether we could observe changes in the penetrance (percentile of microcephalic zebrafish in a clutch) or the expressivity (percentile changes in mean head size) of the *KCTD13*-established phenotype. We observed no changes in penetrance. However, pairwise expression with two other transcripts, *MAPK3* and *MVP* increased significantly the expressivity of the phenotype from 18% for *KCTD13* alone to 24% and 22%, respectively (Supplementary Fig. 6 and Supplementary Table 2), predicting that 16p11.2del/dup patients might have a more severe phenotype than individuals with heterozygous loss of function at *KCTD13* alone.

Finally, we examined whether loss of *KCDT13* in humans might be sufficient to cause some of the commonly observed phenotypes associated with the 16p11.2 deletion. During our analyses, a submicroscopic ~118-kb deletion in 16p11.2 that segregated with ASD and other neurodevelopmental abnormalities was discovered in a single three-generation pedigree[22]. This deletion encompasses five genes, *MVP*, *CDIPT1*, *SEZ6L2*, *ASPHD1* and *KCTD13*, which is consistent with our hypothesis that haploinsufficiency at *KCTD13* might contribute to 16p11.2 phenotypes. We performed a multiplex ligation-dependent probe amplification (MLPA; Supplementary Table 3) assay of this restricted region in 518 subjects that met diagnostic criteria (Autism Diagnostic Observational Schedule, ADOS[10]) for autism or ASD. We found full-segment deletions in eight independent ASD subjects (1.54%; six deletions and two duplications), compared to just five such events from 8,328 controls (0.06%)[23]. We also noted the deletion of a single probe spanning exon 4 of *KCTD13* in one proband with a narrow diagnosis of autism (Supplementary Fig. 7 and Supplementary Table 3). The MLPA assay was replicated in the proband and the deletion was confirmed further by quantitative polymerase chain reaction (qPCR) on two independently obtained DNA samples (Supplementary Fig. 7). Identical MLPA analyses, as well as confirmation of paternity by genotyping, were performed in both biological parents, revealing that the deletion arose *de novo*, and was restricted to the *KCTD13* coding region, maximally spanning 9 kb and including exons 3, 4 and 5 (Supplementary Fig. 7). However, through seeking to precisely localize the breakpoints by custom tiled array comparative genomic hybridization (array CGH) of the entire 16p11.2 region, we found the rearrangement to be more complicated than expected. We discovered an additional, atypical ~300-kb deletion, distal to the classic 16p11.2 region, that was inherited from an asymptomatic mother. The deletion was confirmed by an independent Agilent 24 M feature array CGH, but the precise breakpoints could not be definitely localized as the rearrangement is mediated by a highly complex genomic region of segmental duplication[24]. The apparent complexity of these co-occurring events, including both inherited and *de novo* rearrangements, impacts multiple loci in addition to *KCTD13*, suggesting that this phenotype cannot be attributed solely to the *KCTD13* alteration.

In summary, our data support a major contributory role for *KCDT13* in the 16p11.2 CNV through four lines of evidence: first, the *in vivo* overexpression screen yielded microcephaly in 1 out of 29 genes; second, the reciprocal suppression of this locus mirrored the corresponding human 16p11.2del phenotypes; third, functional analyses established a neurogenic defect that was consistent across species; and fourth, *KCTD13* lies in the putative ~118-kb critical region delineated in one family with ASD independent from our study, and here we found a complex rearrangement that includes a *de novo* alteration affecting a portion of the coding region in a patient with a narrow diagnosis of autism.

Given that the design of our screen relies on a heterologous system of expression, we do not exclude the possibility that other genes in the 16p11.2 CNV might also be relevant to human pathology but did not trigger phenotypes in zebrafish embryos, particularly the five transcripts not present in the zebrafish genome. Moreover, we do not know whether dosage imbalance of *KCTD13* might regulate other phenotypes commonly associated with the 16p11.2 CNV, such as obesity and epilepsy[4–6]. However, it is reasonable to speculate that the neurodevelopmental changes seen after *KCTD13* perturbation could contribute to those phenotypes. We note that loss-of-function mutations of another member of this family, *KCTD7*, cause progressive myoclonic epilepsy[25], and that variants in *KCTD15* have been associated with obesity[26].

KCTD13 encodes the polymerase delta-interacting protein 1 (PDIP1), which interacts with the proliferating cell nuclear antigen[27] and therefore might have a role in the regulation of cell cycle during neurogenesis. Our studies might also shed light on disease architecture and on whether changes in neuronal populations can account for brain overgrowth phenotypes[28]. Although our data suggest that disregulated KCTD13 levels are sufficient to establish neuroanatomical defects, its genetic interaction with at least one more locus leads us to speculate that the point mutation event that would phenocopy the 16p11.2del might not be represented by single loss-of-function alleles at *KCTD13*, but at *cis* or *trans* alleles in *KCTD13*, *MAPK3* or *MVP*, or combined haploinsufficiency of all three. Testing this possibility with sufficient statistical power will require the analysis of large cohorts.

The 16p11.2 CNV has been modelled in mice by chromosomal engineering, and neuroanatomical volumetric changes have been found in regions analogous to those evaluated in our study (for example, in the forebrain and midbrain)[29]. Our approach is likely to

be particularly useful in resolving a rapidly growing number of genomic regions implicated in a range of human genomic disorder phenotypes[30] that are involved in both deletion and duplication syndromes (such as 7q11.23 CNV)[10] and that have mirroring anatomical phenotypes that can be assayed in a physiologically relevant developmental system (such as 1q21.1, and 3q39 CNVs[1]). Such analyses will expedite the identification of major dosage-sensitive loci and accelerate our biological understanding of CNVs that, together, account for a substantial fraction of the mutational burden of neurodevelopmental disorders[1,11].

METHODS SUMMARY

Morpholino, *in vivo* analysis of gene expression and embryo manipulations. The splice-blocking morpholino against Kctd13 was designed and obtained from Gene Tools, LLC (5′-TCTAAGGGTACACGCCTGACCTGTA-3′). Control morpholino was the scrambled nucleotide sequence from Gene Tools, LLC (5′-CCTCTTACCTCAGTTACAATTTATA- 3′). Morpholino injection, RNA rescue and overexpression experiments were performed using a standard protocol (Supplementary Methods).

Immunostaining and TUNEL assay. Embryos were fixed in 4% paraformaldehyde (PFA) or in Dent's fixative (80% methanol, 20% dimethylsulphoxide (DMSO)) overnight at 4 °C. Embryos were incubated in the first antibody solution, 1:750 anti-histone H3 (ser10)-R (sc-8656-R, Santa Cruz), 1:1000 anti-HuC/D (A21271, Invitrogen) and 1:1000 anti-acetylated tubulin (T7451, Sigma-Aldrich). Apoptotic cell death in zebrafish whole-mounts was detected according to a modification of the ApopTag rhodamine *In Situ* Apoptosis Detection kit (Chemicon) protocol. Embryos for cryosectioning were fixed, incubated in PBS containing sucrose 30%, embedded in Tissue-Tek O.C.T. Embedding Compound (Sakura Finetek) and sectioned at 7 μm.

***In utero* electroporation.** E13.5 embryos were injected with either control or *Kctd13* shRNA constructs and a GFP-expressing vector. Forty-eight hours after electroporation, BrdU (100 mg kg^{-1}) was injected intraperitoneally into the pregnant dams and embryos were harvested 2 h later. Embryo brains were processed and sectioned (20 μm) before staining with a BrdU antibody (Accurate).

Short arm of chromosome 16 custom array CGH. DNA samples from the proband and both parents were labelled with Cy3 and co-hybridized with Cy5-labelled control DNA from an unaffected CEPH (Centre d'Étude du Polymorphisme Humain) individual obtained from Coriell (GM10851) to custom-made Nimblegen arrays. DNA labelling, hybridization and washing were performed according to Nimblegen protocols. Scanning was performed using a Roche/ Nimblegen MS200 Scanner. Image processing, quality control and data extraction were performed using the Nimblescan software v.2.6.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 4 September 2011; accepted 27 March 2012.

1. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61,** 437–455 (2010).
2. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358,** 667–675 (2008).
3. McCarthy, S. E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nature Genet.* **41,** 1223–1227 (2009).
4. Jacquemont, S. *et al.* Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* **478,** 97–102 (2011).
5. Walters, R. G. *et al.* A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* **463,** 671–675 (2010).
6. Kumar, R. A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.* **17,** 628–638 (2008).
7. Baldini, A. Dissecting contiguous gene defects: TBX1. *Curr. Opin. Genet. Dev.* **15,** 279–284 (2005).
8. Lupski, J. R. *et al.* DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* **66,** 219–232 (1991).
9. Brunetti-Pierri, N. *et al.* Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nature Genet.* **40,** 1466–1471 (2008).
10. Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70,** 863–885 (2011).
11. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316,** 445–449 (2007).
12. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455,** 232–236 (2008).
13. Shinawi, M. *et al.* Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. *J. Med. Genet.* **47,** 332–341 (2010).
14. Roa, B. B. *et al.* Charcot-Marie-Tooth disease type 1A. Association with a spontaneous point mutation in the PMP22 gene. *N. Engl. J. Med.* **329,** 96–101 (1993).
15. Slager, R. E., Newton, T. L., Vlangos, C. N., Finucane, B. & Elsea, S. H. Mutations in RAI1 associated with Smith-Magenis syndrome. *Nature Genet.* **33,** 466–468 (2003).
16. Talkowski, M. E. *et al.* Assessment of 2q23.1 microdeletion syndrome implicates MBD5 as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder. *Am. J. Hum. Genet.* **89,** 551–563 (2011).
17. Lindsay, E. A. *et al.* Tbx1 haploinsufficieny in the DiGeorge syndrome region causes aortic arch defects in mice. *Nature* **410,** 97–101 (2001).
18. Peça, J. *et al.* Shank3 mutant mice display autistic-like behaviours and striatal dysfunction. *Nature* **472,** 437–442 (2011).
19. Ceol, C. J. *et al.* The histone methyltransferase SETDB1 is recurrently amplified in melanoma and accelerates its onset. *Nature* **471,** 513–517 (2011).
20. Takeda, H., Matsuzaki, T., Oki, T., Miyagawa, T. & Amanuma, H. A novel POU domain gene, zebrafish pou2: expression and roles of two alternatively spliced twin products in early development. *Genes Dev.* **8,** 45–59 (1994).
21. Detrich, H. W. III, Westerfield, M. & Zon, L. I. Overview of the Zebrafish system. *Methods Cell Biol.* **59,** 3–10 (1999).
22. Crepel, A. *et al.* Narrowing the critical deletion region for autism spectrum disorders on 16p11.2. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **156,** 243–245 (2011).
23. Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nature Genet.* **43,** 838–846 (2011).
24. Barge-Schaapveld, D. Q., Maas, S. M., Polstra, A., Knegt, L. C. & Hennekam, R. C. The atypical 16p11.2 deletion: a not so atypical microdeletion syndrome? *Am. J. Med. Genet. A.* **155,** 1066–1072 (2011).
25. Azizieh, R. *et al.* Progressive myoclonic epilepsy-associated gene KCTD7 is a regulator of potassium conductance in neurons. *Mol. Neurobiol.* **44,** 111–121 (2011).
26. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.* **41,** 25–34 (2009).
27. He, H., Tan, C. K., Downey, K. M. & So, A. G. A tumor necrosis factor alpha- and interleukin 6-inducible protein that interacts with the small subunit of DNA polymerase delta and proliferating cell nuclear antigen. *Proc. Natl Acad. Sci. USA* **98,** 11979–11984 (2001).
28. Courchesne, E. *et al.* Neuron number and size in prefrontal cortex of children with autism. *J. Am. Med. Assoc.* **306,** 2001–2010 (2011).
29. Horev, G. *et al.* Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proc. Natl Acad. Sci. USA* **108,** 17076–17081 (2011).
30. Beckmann, J. S., Estivill, X. & Antonarakis, S. E. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nature Rev. Genet.* **8,** 639–646 (2007).
31. Crespi, B., Stead, P. & Elliot, M. Evolution in health and medicine Sackler colloquium: comparative genomics of autism and schizophrenia. *Proc. Natl Acad. Sci. USA* **107** (suppl. 1), 1736–1741 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Hanscom for technical assistance. We also thank J. Black, J. Whetstine, K. Brown and C. Lee for assistance with the initial and confirmatory array CGH experiments, S. Santagelo for analysis of clinical phenotype data, and M. State, M. Daley and R. Gibbs for sharing unpublished exome sequencing data. This work was supported by a Silvo O. Conte center grant (MH-094268) from the National Institute of Mental Health (NIMH), National Institutes of Health grant MH-084018 (A.S., A.K. and N.K.), grant MH-091230 (A.K.), grant HD06286 (J.F.G.), the Simon's Foundation, the Autism Consortium of Boston (J.F.G.), the Leenaards Foundation Prize (S.J. and A.R.), the Swiss National Science Foundation (A.R. and J.S.B.), a Swiss National Science Foundation Sinergia grant (S.J., J.S.B. and A.R.). M.E.T. was supported by an NIMH National Research Service Award (F32MH087123). S.J. is a recipient of a bourse de relève académique de la Faculté de Biologie et Médecine de l'Université de Lausanne. N.K. is a Distinguished Brumley Professor.

Author Contributions C.G. and N.K. designed the study and wrote the paper, and all authors approved and commented on the manuscript. C.G. performed the zebrafish studies, immunostaining, TUNEL and sectioning experiments, and counting analyses. J.W. made the plasmid constructs and capped mRNAs. M.E.T., M.S. and J.F.G. performed the human genetic analyses. E.C.O. performed the shRNA silencing and BrdU pulse experiments, and generated mice brain sections. Y.T., A.S. and A.K. performed *in utero* electroporations. S.J., A.R. and J.S.B. designed the 16p-specific microarray and shared unpublished data that informed our experimental design.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.K. (katsanis@cellbio.duke.edu).

METHODS

Morpholino, *in vivo* analysis of gene expression and embryo manipulations. The splice-blocking morpholino against Kctd13 was designed and obtained from Gene Tools (5′-TCTAAGGGTACACGCCTGACCTGTA-3′). Control morpholino was the scrambled nucleotide sequence from Gene Tools (5′-CCTCTTAC CTCAGTTACAATTTATA-3′). We injected 1 nl of diluted morpholino (6, 8 or 10 ng) and/or RNA (50, 75 or 100 pg) into wild-type zebrafish embryos at the 1- to 2-cell stage. Injected embryos were scored at 4.25 d.p.f. and classified into two groups, normal and mutant, on the basis of the relative head size compared with age-matched controls from the same clutch. For RNA rescue and overexpression experiments, the human wild-type mRNAs were cloned into the pCS2 vector and transcribed *in vitro* using the SP6 Message Machine kit (Ambion). All the experiments were repeated three times and we ran a *t*-test to determine the significance of the morphant phenotype. Whole-mount *in situ* hybridizations were carried out using antisense probes for *kctd13* made from clone (Openbiosystems) and following the manufacturer's protocols.

Whole-mount TUNEL assay. Apoptotic cell death in zebrafish whole-mounts was detected according to a modification of the ApopTag rhodamine *In Situ* Apoptosis Detection kit (Chemicon) protocol. Embryos were fixed in 4% PFA at 4 °C overnight and store in 100% methanol at −20 °C. After rehydratation in PBS, embryos were permeabilized by a 5-min digestion with proteinase K (10 μg ml^{-1}) in PBS at room temperature. After two washes in sterile water for 3 min each, embryos were postfixed with 4% PFA for 20 min at room temperature (20 °C–22 °C) and then with prechilled ethanol:acetic acid (2:1) for 10 min at −20 °C. Embryos were washed in PBS-T (PBS 1×, 0.1% Tween 20) for 5 min, three times at room temperature. The incubation in the equilibration buffer and further steps were followed according to the standard protocol suggested by the manufacturer. TUNEL staining was quantified by counting positive cells in defined regions of the head, and using ImageJ software.

Zebrafish whole-mount and section immunostaining. Embryos were fixed in 4% PFA overnight and stored in 100% methanol at −20 °C. For acetylated tubulin staining, embryos were fixed in Dent's fixative (80% methanol, 20% dimethylsulphoxide (DMSO)) overnight at 4 °C. The embryos were permeabilized with proteinase K, then postfixed with 4% PFA, washed in PBSTX (PBS+0.5%, Triton X-100). After rehydration in PBS, PFA-fixed embryos were washed in IF buffer (0.1% Tween-20, 1% BSA in PBS 1×) for 10 min at room temperature. The embryos were incubated in the blocking buffer (10% FBS, 1% BSA in PBS 1×) for 1 h at room temperature. After two washes in IF Buffer for 10 min each, embryos were incubated in the first antibody solution, 1:750 anti-histone H3 (ser10)-R, (sc-8656-R, Santa Cruz), 1:1000 anti-HuC/D (A21271, Invitrogen), 1:1000 anti-acetylated tubulin (T7451, Sigma-Aldrich), in blocking solution, overnight at 4 °C. After two washes in IF Buffer for 10 min each, embryos were incubated in the secondary antibody solution, 1:1000 Alexa Fluor donkey anti-rabbit IgG and Alexa Fluor goat anti-mouse IgG (A21207, A11001, Invitrogen), in blocking solution, for 1 h at room temperature. Staining was quantified by counting positive cells in defined regions of the head and using ImageJ software. Embryos for cryosectioning were fixed in 4% PFA overnight at 4 °C, washed twice in PBS and transferred to 30% sucrose in PBS at 4 °C overnight. Embryos were embedded in Tissue-Tek O.C.T. Embedding Compound (Sakura Finetek), and sections were cut (7 μm thick).

Cell-proliferation assay. Neuro-2a cells were seeded in two-well chamber slides and transfected with control and *Kctd13* shRNA constructs and a GFP-expressing plasmid. Two days after transfection, cells were pulsed with 10 μM BrdU, then fixed and stained with a BrdU antibody (Accurate).

***In utero* electroporation.** E13.5 embryos were injected with either control or *Kctd13* shRNA constructs and a GFP-expressing vector. Forty-eight hours after electroporation, BrdU (100 mg kg^{-1}) was injected intraperitoneally into the pregnant dams and embryos were harvested 2 h later. Embryo brains were processed and sectioned (20 μm) before staining with a BrdU antibody (Accurate).

MLPA. A custom assay with 29 probes was designed to capture candidate genes in the 118-kb 16p11.2 putative critical region (23 probes to the targeted genes *KCTD13*, *MVP* and *CDIPT*; 2 probes to 16p11.2 microdeletion region genes not within the putative critical region, *TAOK2* and *TBX6*; 4 probes to control genes outside of the region). The assay was performed with 100–150 ng genomic DNA (quantified byPico-Green (Quant-iT, Invitrogen)) according to the manufacturer's instructions. All samples were performed in triplicate. Amplification products from ligated probes were run on an ABI 3730xl DNA Analyzer using Genescan-Rox500 size standards (Applied Biosystems). The raw 3730xl electrophoresis signals from MLPA probe amplification were processed using GeneMarker Software Trial Version 1.91 (SoftGenetics) with the size standards as a reference, and a ratio of peak heights and areas from each electrophoresis signal were compared between control individuals and ASD samples. Ratio values between 0.75 and 1.3 were considered normal.

RT–PCR. Each quantitative real-time PCR was performed in a 96-well plate using 2 x LightCycler 480 SYBR Green I Master mix (Roche Applied Science) according to the manufacturer's instructions. All products were cycled at 95 °C for 2 min, then by 45–55 cycles of 95 °C for 15 s and 60 °C for 1 min. Melt-temperature analysis was performed at the end of each run to confirm PCR specificity. Relative copy number was determined between the proband (AC02-1467-01) and three control subjects for probes designed within exon 4 of KCTD13 and control probes localized outside of the 16p11.2 microdeletion region.

Short arm of chromosome 16 custom array CGH. DNA samples from the proband and both parents were labelled with Cy3 and co-hybridized with Cy5-labelled control DNA from an unaffected CEPH individual, obtained from Coriell (GM10851), to custom-made Nimblegen arrays. These arrays contained 71,000 probes spread across the short arm of chromosome 16, from 22.0 to 32.7 Mb (at a median space of 45 bp between 27.5 and 31.0 Mb) and 1,000 control probes situated in invariable region of the X chromosome. DNA labelling, hybridization and washing were performed according to Nimblegen protocols. Scanning was performed using a Roche/Nimblegen MS200 Scanner. Image processing, quality control and data extraction were performed using the Nimblescan software v.2.6.

LETTER

doi:10.1038/nature11082

Topological domains in mammalian genomes identified by analysis of chromatin interactions

Jesse R. Dixon[1,2,3], Siddarth Selvaraj[1,4], Feng Yue[1], Audrey Kim[1], Yan Li[1], Yin Shen[1], Ming Hu[5], Jun S. Liu[5] & Bing Ren[1,6]

The spatial organization of the genome is intimately linked to its biological function, yet our understanding of higher order genomic structure is coarse, fragmented and incomplete. In the nucleus of eukaryotic cells, interphase chromosomes occupy distinct chromosome territories, and numerous models have been proposed for how chromosomes fold within chromosome territories[1]. These models, however, provide only few mechanistic details about the relationship between higher order chromatin structure and genome function. Recent advances in genomic technologies have led to rapid advances in the study of three-dimensional genome organization. In particular, Hi-C has been introduced as a method for identifying higher order chromatin interactions genome wide[2]. Here we investigate the three-dimensional organization of the human and mouse genomes in embryonic stem cells and terminally differentiated cell types at unprecedented resolution. We identify large, megabase-sized local chromatin interaction domains, which we term 'topological domains', as a pervasive structural feature of the genome organization. These domains correlate with regions of the genome that constrain the spread of heterochromatin. The domains are stable across different cell types and highly conserved across species, indicating that topological domains are an inherent property of mammalian genomes. Finally, we find that the boundaries of topological domains are enriched for the insulator binding protein CTCF, housekeeping genes, transfer RNAs and short interspersed element (SINE) retrotransposons, indicating that these factors may have a role in establishing the topological domain structure of the genome.

To study chromatin structure in mammalian cells, we determined genome-wide chromatin interaction frequencies by performing the Hi-C experiment[2] in mouse embryonic stem (ES) cells, human ES cells, and human IMR90 fibroblasts. Together with Hi-C data for the mouse cortex generated in a separate study (Y. Shen *et al.*, manuscript in preparation), we analysed over 1.7-billion read pairs of Hi-C data corresponding to pluripotent and differentiated cells (Supplementary Table 1). We normalized the Hi-C interactions for biases in the data (Supplementary Figs 1 and 2)[3]. To validate the quality of our Hi-C data, we compared the data with previous chromosome conformation capture (3C), chromosome conformation capture carbon copy (5C), and fluorescence *in situ* hybridization (FISH) results[4–6]. Our IMR90 Hi-C data show a high degree of similarity when compared to a previously generated 5C data set from lung fibroblasts (Supplementary Fig. 4). In addition, our mouse ES cell Hi-C data correctly recovered a previously described cell-type-specific interaction at the *Phc1* gene[5] (Supplementary Fig. 5). Furthermore, the Hi-C interaction frequencies in mouse ES cells are well-correlated with the mean spatial distance separating six loci as measured by two-dimensional FISH[6] (Supplementary Fig. 6), demonstrating that the normalized Hi-C data can accurately reproduce the expected nuclear distance using an independent method. These results demonstrate that our Hi-C data are of high quality and accurately capture the higher order chromatin structures in mammalian cells.

We next visualized two-dimensional interaction matrices using a variety of bin sizes to identify interaction patterns revealed as a result of our high sequencing depth (Supplementary Fig. 7). We noticed that at bin sizes less than 100 kilobases (kb), highly self-interacting regions begin to emerge (Fig. 1a and Supplementary Fig. 7, seen as 'triangles' on the heat map). These regions, which we term topological domains, are bounded by narrow segments where the chromatin interactions appear to end abruptly. We hypothesized that these abrupt transitions may represent boundary regions in the genome that separate topological domains.

To identify systematically all such topological domains in the genome, we devised a simple statistic termed the directionality index to quantify the degree of upstream or downstream interaction bias for a genomic region, which varies considerably at the periphery of the topological domains (Fig. 1b; see Supplementary Methods for details). The directionality index was reproducible (Supplementary Table 2) and pervasive, with 52% of the genome having a directionality index that was not expected by random chance (Fig. 1c, false discovery rate = 1%). We then used a Hidden Markov model (HMM) based on the directionality index to identify biased 'states' and therefore infer the locations of topological domains in the genome (Fig. 1a; see Supplementary Methods for details). The domains defined by HMM were reproducible between replicates (Supplementary Fig. 8). Therefore, we combined the data from the HindIII replicates and identified 2,200 topological domains in mouse ES cells with a median size of 880 kb that occupy ~91% of the genome (Supplementary Fig. 9). As expected, the frequency of intra-domain interactions is higher than inter-domain interactions (Fig. 1d, e). Similarly, FISH probes[6] in the same topological domain (Fig. 1f) are closer in nuclear space than probes in different topological domains (Fig. 1g), despite similar genomic distances between probe pairs (Fig. 1h, i). These findings are best explained by a model of the organization of genomic DNA into spatial modules linked by short chromatin segments. We define the genomic regions between topological domains as either 'topological boundary regions' or 'unorganized chromatin', depending on their sizes (Supplementary Fig. 9).

We next investigated the relationship between the topological domains and the transcriptional control process. The *Hoxa* locus is separated into two compartments by an experimentally validated insulator[4,7,8], which we observed corresponds to a topological domain boundary in both mouse (Fig. 1a) and human (Fig. 2a). Therefore, we hypothesized that the boundaries of the topological domains might correspond to insulator or barrier elements.

Many known insulator or barrier elements are bound by the zinc-finger-containing protein CTCF (refs 9–11). We see a strong enrichment of CTCF at the topological boundary regions (Fig. 2b and Supplementary Fig. 10), indicating that topological boundary regions

[1]Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, California 92093, USA. [2]Medical Scientist Training Program, University of California, San Diego, La Jolla, California 92093, USA. [3]Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla, California 92093, USA. [4]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, California 92093, USA. [5]Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, Massachusetts 02138, USA. [6]University of California, San Diego School of Medicine, Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, UCSD Moores Cancer Center, 9500 Gilman Drive, La Jolla, California 92093, USA.

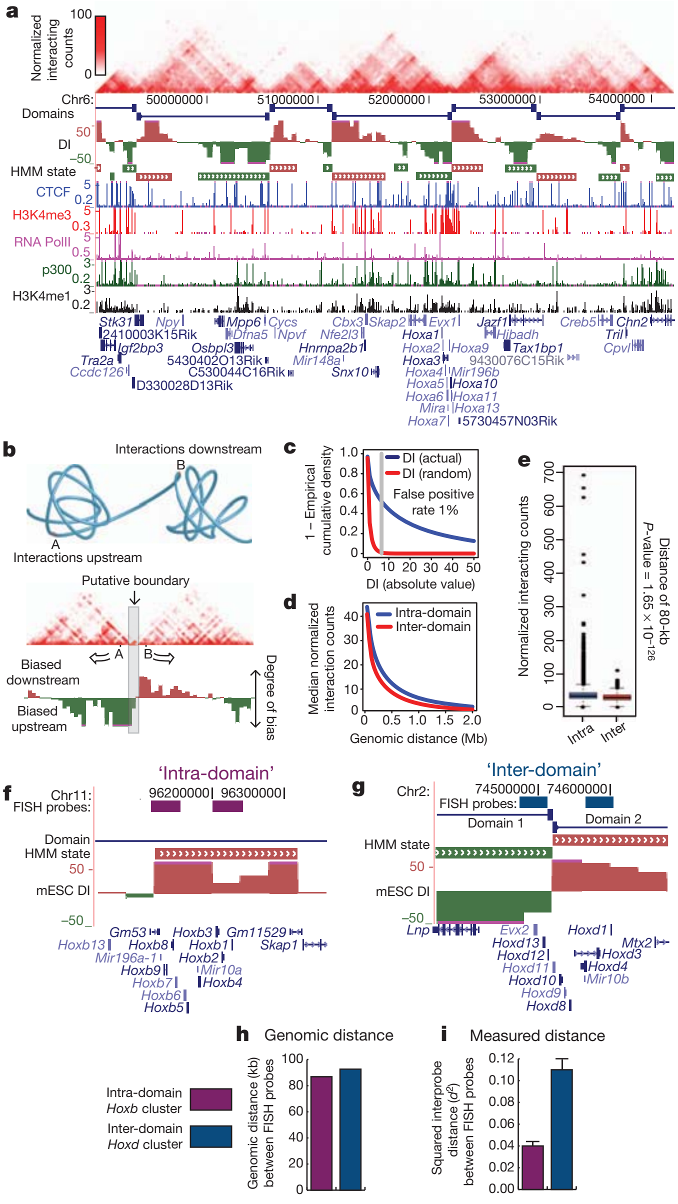


Figure 1 | Topological domains in the mouse ES cell genome. a, Normalized Hi-C interaction frequencies displayed as a two-dimensional heat map overlayed on ChIP-seq data (from Y. Shen *et al.*, manuscript in preparation), directionality index (DI), HMM bias state calls, and domains. For both directionality index and HMM state calls, downstream bias (red) and upstream bias (green) are indicated. **b**, Schematic illustrating topological domains and resulting directional bias. **c**, Distribution of the directionality index (absolute value, in blue) compared to random (red). **d**, Mean interaction frequencies at all genomic distances between 40 kb to 2 Mb. Above 40 kb, the intra- versus inter-domain interaction frequencies are significantly different ($P < 0.005$, Wilcoxon test). **e**, Box plot of all interaction frequencies at 80-kb distance. Intra-domain interactions are enriched for high-frequency interactions. **f**–**i**, Diagram of intra-domain (**f**) and inter-domain FISH probes (**g**) and the genomic distance between pairs (**h**). **i**, Bar chart of the squared inter-probe distance (from ref. 6) FISH probe pairs. mESC, mouse ES cell. Error bars indicate standard error ($n = 100$ for each probe pair).

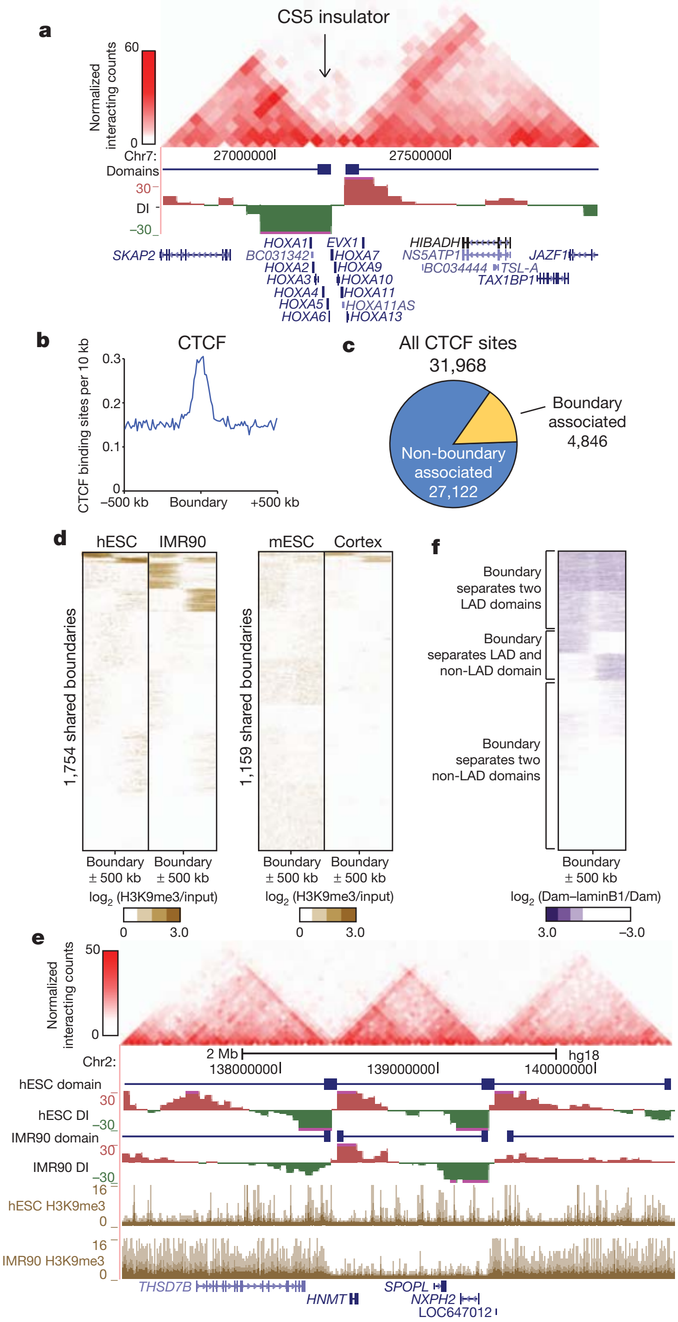


Figure 2 | Topological boundaries demonstrate classical insulator or barrier element features. a, Two-dimensional heat map surrounding the *Hoxa* locus and CS5 insulator in IMR90 cells. **b**, Enrichment of CTCF at boundary regions. **c**, The portion of CTCF binding sites that are considered 'associated' with a boundary (within ±20-kb window is used as the expected uncertainty due to 40-kb binning). **d**, Heat maps of H3K9me3 at boundary sites in human and mouse. **e**, UCSC Genome Browser shot showing heterochromatin spreading in the human ES cells (hESC) and IMR90 cells. The two-dimensional heat map shows the interaction frequency in human ES cells. **f**, Heat map of LADs (from ref. 14) surrounding the boundary regions. Scale is the $\log_2$ ratio of DNA adenosine methylation (Dam)–lamin B1 fusion over Dam alone (Dam–laminB1/Dam).

share this feature of classical insulators. A classical boundary element is also known to stop the spread of heterochromatin. Therefore, we examined the distribution of the heterochromatin mark H3K9me3 in humans and mice in relation to the topological domains[12,13]. Indeed, we observe a clear segregation of H3K9me3 at the boundary regions that occurs predominately in differentiated cells (Fig. 2d, e and Supplementary Fig. 11). As the boundaries that we analysed in Fig. 2d are present in both pluripotent cells and their differentiated progeny, the topological domains and boundaries appear to pre-mark the end points of heterochromatic spreading. Therefore, the domains do not seem to be a consequence of the formation of heterochromatin. Taken together, the above observations strongly suggest that the topological domain boundaries correlate with regions of the genome displaying classical insulator and barrier element activity, thus revealing a

potential link between the topological domains and transcriptional control in the mammalian genome.

We compared the topological domains with previously described domain-like organizations of the genome, specifically with the A and B compartments described by ref. 2, with lamina-associated domains (LADs)[10,14], replication time zones[15,16], and large organized chromatin K9 modification (LOCK) domains[17]. In all cases, we can see that topological domains are related to, but independent from, each of these previously described domain-like structures (Supplementary Figs 12–15). Notably, a subset of the domain boundaries we identify appear to mark the transition between either LAD and non-LAD regions of the genome (Fig. 2f and Supplementary Fig. 12), the A and B compartments (Supplementary Fig. 13, 14), and early and late replicating chromatin (Supplementary Fig. 14). Lastly, we can also confirm the previously reported similarities between the A and B compartments and early and late replication time zone (Supplementary Fig. 16)[16].

We next compared the locations of topological boundaries identified in both replicates of mouse ES cells and cortex, or between both replicates of human ES cells and IMR90 cells. In both human and mouse, most of the boundary regions are shared between cell types (Fig. 3a and Supplementary Fig. 17a), suggesting that the overall domain structure between cell types is largely unchanged. At the boundaries called in only one cell type, we noticed that trend of upstream and downstream bias in the directionality index is still readily apparent and highly reproducible between replicates (Supplementary Fig. 17b, c). We cannot determine if the differences in domain calls between cell types is due to noise in the data or to biological phenomena, such as a change in the strength of the boundary region between cell types[18]. Regardless, these results indicate that the domain boundaries are largely invariant between cell types. Lastly, only a small fraction of the boundaries show clear differences between two cell types, suggesting that a relatively rare subset of boundaries may actually differ between cell types (Supplementary Fig. 18).

The stability of the domains between cell types is surprising given previous evidence showing cell-type-specific chromatin interactions and conformations[5,7]. To reconcile these results, we identified cell-type-specific chromatin interactions between mouse ES cell and mouse cortex. We identified 9,888 dynamic interacting regions in the mouse genome based on 20-kb binning using a binomial test with an empirical false discovery rate of <1% based on random permutation of the replicate data. These dynamic interacting regions are enriched for differentially expressed genes (Fig. 3b–d, Supplementary Fig. 19 and Supplementary Table 5). In fact, 20% of all genes that undergo a four-fold change in gene expression are found at dynamic interacting loci. This is probably an underestimate, because by binning the genome at 20 kb, any dynamic regulatory interaction less than 20 kb will be missed. Lastly, >96% of dynamic interacting regions occur in the same domain (Fig. 3e). Therefore, we favour a model where the domain organization is stable between cell types, but the regions within each domain may be dynamic, potentially taking part in cell-type-specific regulatory events.

The stability of the domains between cell types prompted us to investigate if the domain structure is also conserved across evolution. To address this, we compared the domain boundaries between mouse ES cells and human ES cells using the UCSC liftover tool. Most of the boundaries appear to be shared across evolution (53.8% of human boundaries are boundaries in mouse and 75.9% of mouse boundaries are boundaries in humans, compared to 21.0% and 29.0% at random, P value $<2.2 \times 10^{-16}$, Fisher's exact test; Fig. 3f). The syntenic regions in mouse and human in particular share a high degree of similarity in their higher order chromatin structure (Fig. 3g, h), indicating that there is conservation of genomic structure beyond the primary sequence of DNA.

We explored what factors may contribute to the formation of topological boundary regions in the genome. Although most topological boundaries are enriched for the binding of CTCF, only 15% of CTCF

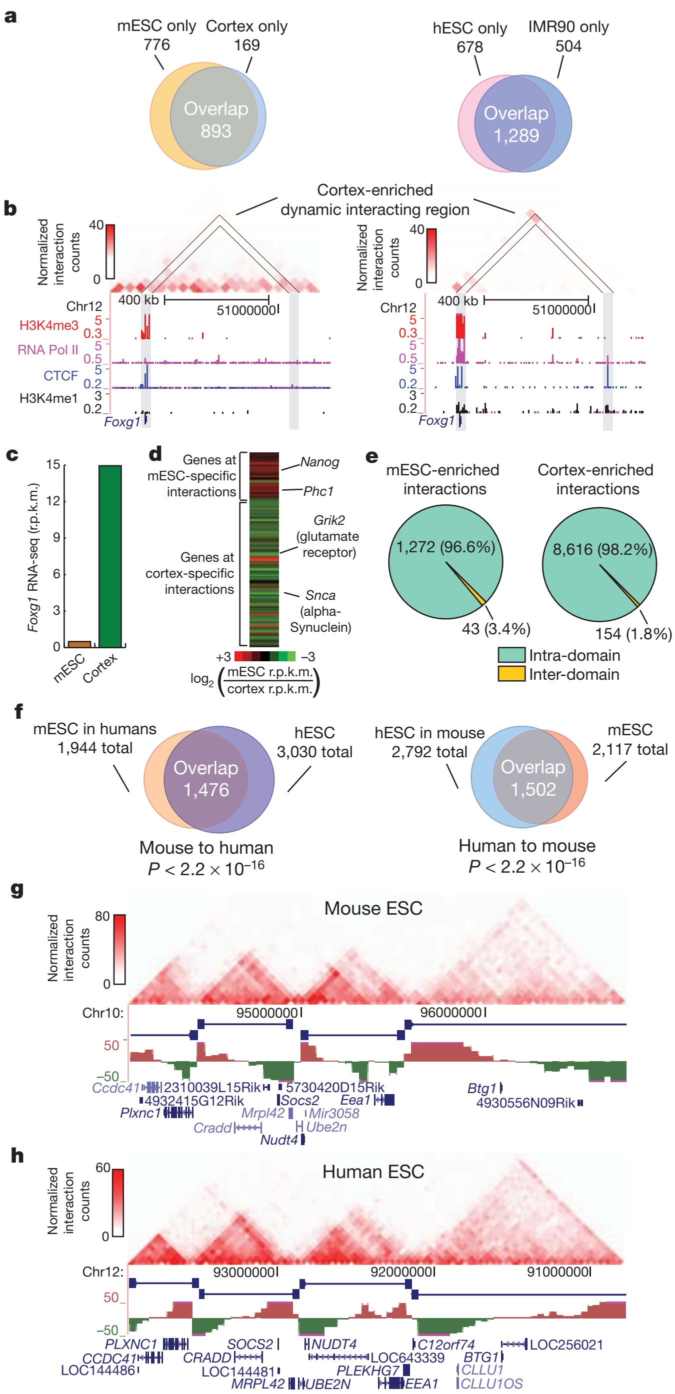


Figure 3 | Boundaries are shared across cell types and conserved in evolution. **a**, Overlap of boundaries between cell types. **b**, Genome browser shot of a cortex enriched dynamic interacting region that overlaps with the *Foxg1* gene. **c**, *Foxg1* expression in reads per kilobase per million reads sequenced (r.p.k.m.) in mouse ES cells and cortex as measured by RNA-seq. **d**, Heat map of the gene expression ratio between mouse ES cell and cortex of genes at dynamic interactions. **e**, Pie chart of inter- and intra-domain dynamic interactions. **f**, Overlap of boundaries between syntenic mouse and human sequences ($P < 2.2 \times 10^{-16}$ compared to random, Fisher's exact test). **g**, **h**, Genome browser shots showing domain structure over a syntenic region in the mouse (**g**) and human (**h**) ES cells. Note: the region in humans has been inverted from its normal UCSC coordinates for proper display purposes.

binding sites are located within boundary regions (Fig. 2c). Thus, CTCF binding alone is insufficient to demarcate domain boundaries. We reasoned that additional factors might be associated with topological boundary regions. By examining the enrichment of a variety of

histone modifications, chromatin binding proteins and transcription factors around topological boundary regions in mouse ES cells, we observed that factors associated with active promoters and gene bodies are enriched at boundaries in both mouse and humans (Fig. 4a and Supplementary Figs 20–23)[19,20]. In contrast, non-promoter-associated marks, such as H3K4me1 (associated with enhancers) and H3K9me3, were not enriched or were specifically depleted at boundary regions (Fig. 4a). Furthermore, transcription start sites (TSS) and global run on sequencing (GRO-seq)[21] signal were also enriched around topological boundaries (Fig. 4a). We found that housekeeping genes were particularly strongly enriched near topological boundary regions (Fig. 4b–d; see Supplementary Table 7 for complete GO terms enrichment). Additionally, the tRNA genes, which have the potential to function as boundary elements[22,23], are also enriched at boundaries (P value <0.05, Fisher's exact test; Fig. 4b). These results suggest that high levels of transcription activity may also contribute to boundary formation. In support of this, we can see examples of dynamic changes in H3K4me3 at or near some cell-type-specific boundaries that are cell-type specific (Supplementary Fig. 24). Indeed, boundaries associated with both CTCF and a housekeeping gene account for nearly one-third of all topological boundaries in the genome (Fig. 4e and Supplementary Fig. 24).

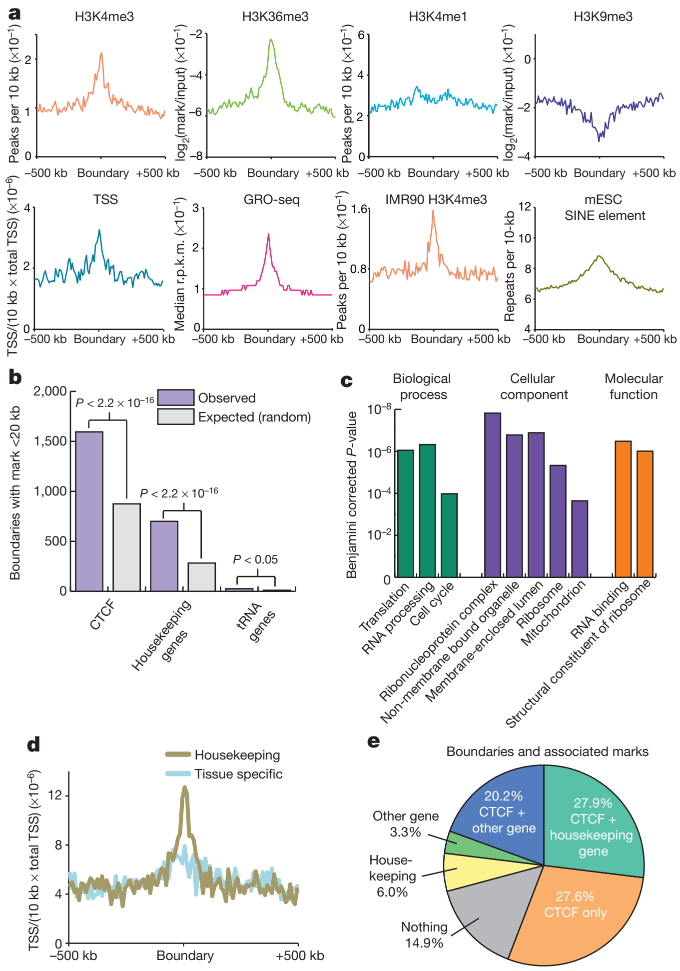


Figure 4 | Boundary regions are enriched for housekeeping genes. **a**, Chromatin modifications, TSS, GRO-seq and SINE elements surrounding boundary regions in mouse ES cells or IMR90 cells. **b**, Boundaries associated with a CTCF binding site, housekeeping gene, or tRNA gene (purple) compared to expected at random (grey). **c**, Gene Ontology P-value chart. **d**, Enrichment of housekeeping genes (gold) and tissue-specific genes (blue) as defined by Shannon entropy scores near boundaries normalized for the number of genes in each class (TSS/10 kb/total TSS). **e**, Percentage of boundaries with a given mark within 20 kb of the boundaries.

Finally, we analysed the enrichment of repeat classes around boundary elements. We observed that Alu/B1 and B2 SINE retrotransposons in mouse and Alu SINE elements in humans are enriched at boundary regions (Fig. 4a and Supplementary Figs 24 and 25). In light of recent reports indicating that a SINE B2 element functions as a boundary in mice[24], and SINE element retrotransposition may alter CTCF binding sites during evolution[25], we believe that this contributes to a growing body of evidence indicating a role for SINE elements in the organization of the genome.

In summary, we show that the mammalian chromosomes are segmented into megabase-sized topological domains, consistent with some previous models of the higher order chromatin structure[1,26,27]. Such spatial organization seems to be a general property of the genome: it is pervasive throughout the genome, stable across different cell types and highly conserved between mice and humans.

We have identified multiple factors that are associated with the boundary regions separating topological domains, including the insulator binding factor CTCF, housekeeping genes and SINE elements. The association of housekeeping genes with boundary regions extends previous studies in yeast and insects and suggests that non-CTCF factors may also be involved in insulator/barrier functions in mammalian cells[28].

The topological domains we identified are well conserved between mice and humans. This indicates that the sequence elements and mechanisms that are responsible for establishing higher order structures in the genome may be relatively ancient in evolution. A similar partitioning of the genome into physical domains has also been observed in *Drosophila* embryos[29] and in high-resolution studies of the X-inactivation centre in mice (termed topologically associated domains or TADs)[30], indicating that topological domains may be a fundamental organizing principle of metazoan genomes.

METHODS SUMMARY

Cell culture and Hi-C experiments. J1 mouse ES cells were grown on gamma-irradiated mouse embryonic fibroblasts cells under standard conditions (85% high glucose DMEM, 15% HyClone FBS, 0.1 mM non-essential amino acids, 0.1 mM β-mercaptoethanol, 1 mM glutamine, LIF 500 U ml^{-1}, 1× Gibco penicillin/streptomycin). Before collecting for Hi-C, J1 mouse ES cells were passaged onto feeder free 0.2% gelatin-coated plates for at least two passages to rid the culture of feeder cells. H1 human ES cells and IMR90 fibroblasts were grown as previously described[13]. Collecting the cells for Hi-C was performed as previously described, with the only modification being that the adherent cell cultures were dissociated with trypsin before fixation.

Sequencing and mapping of data. Hi-C analysis and paired-end libraries were prepared as previously described[2] and sequenced on the Illumina Hi-Seq2000 platform. Reads were mapped to reference human (hg18) or mouse genomes (mm9), and non-mapping reads and PCR duplicates were removed. Two-dimensional heat maps were generated as previously described[2].

Data analysis. For detailed descriptions of the data analysis, including descriptions of the directionality index, hidden Markov models, dynamic interactions identification, and boundary overlap between cells and across species, see Supplementary Methods.

Received 26 September 2011; accepted 27 March 2012.
Published online 11 April 2012.

1. Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harb. Perspect. Biol.* **2,** a003889 (2010).
2. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
3. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genet.* **43,** 1059–1065 (2011).
4. Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472,** 120–124 (2011).
5. Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467,** 430–435 (2010).
6. Eskeland, R. *et al.* Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol. Cell* **38,** 452–464 (2010).
7. Noordermeer, D. *et al.* The dynamic architecture of Hox gene clusters. *Science* **334,** 222–225 (2011).

8. Kim, Y. J., Cecchini, K. R. & Kim, T. H. Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. *Proc. Natl Acad. Sci. USA* **108,** 7391–7396 (2011).
9. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137,** 1194–1211 (2009).
10. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).
11. Handoko, L. *et al.* CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature Genet.* **43,** 630–638 (2011).
12. Xie, W. *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148,** 816–831 (2012).
13. Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6,** 479–491 (2010).
14. Peric-Hupkes, D. *et al.* Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol. Cell* **38,** 603–613 (2010).
15. Hiratani, I. *et al.* Genome-wide dynamics of replication timing revealed by *in vitro* models of mouse embryogenesis. *Genome Res.* **20,** 155–169 (2010).
16. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20,** 761–770 (2010).
17. Wen, B., Wu, H., Shinkai, Y., Irizarry, R. A. & Feinberg, A. P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nature Genet.* **41,** 246–250 (2009).
18. Scott, K. C., Taubman, A. D. & Geyer, P. K. Enhancer blocking by the *Drosophila* gypsy insulator depends upon insulator anatomy and enhancer strength. *Genetics* **153,** 787–798 (1999).
19. Bilodeau, S., Kagey, M. H., Frampton, G. M., Rahl, P. B. & Young, R. A. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev.* **23,** 2484–2489 (2009).
20. Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134,** 521–533 (2008).
21. Min, I. M. *et al.* Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev.* **25,** 742–754 (2011).
22. Donze, D. & Kamakaka, R. T. RNA polymerase III and RNA polymerase II promoter complexes are heterochromatin barriers in *Saccharomyces cerevisiae. EMBO J.* **20,** 520–531 (2001).
23. Ebersole, T. *et al.* tRNA genes protect a reporter gene from epigenetic silencing in mouse cells. *Cell Cycle* **10,** 2779–2791 (2011).
24. Lunyak, V. V. *et al.* Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* **317,** 248–251 (2007).
25. Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell.* **148,** 335–348 (2012).
26. Jhunjhunwala, S. *et al.* The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell* **133,** 265–279 (2008).
27. Capelson, M. & Corces, V. G. Boundary elements and nuclear organization. *Biol. Cell* **96,** 617–629 (2004).
28. Amouyal, M. Gene insulation. Part I: natural strategies in yeast and *Drosophila. Biochem. Cell Biol.* **88,** 875–884 (2010).
29. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148,** 458–472 (2012).
30. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* http://dx.doi.org/10.1038/nature11049 (this issue).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful for the comments from and discussions with Z. Qin, A. Desai and members of the Ren laboratory during the course of the study. We also thank W. Bickmore and R. Eskeland for sharing the FISH data generated in mouse ES cells. This work was supported by funding from the Ludwig Institute for Cancer Research, California Institute for Regenerative Medicine (CIRM, RN2-00905-1) (to B.R.) and NIH (B.R. R01GH003991). J.R.D. is funded by a pre-doctoral training grant from CIRM. Y.S. is supported by a postdoctoral fellowship from the Rett Syndrome Research Foundation.

Author Contributions J.R.D. and B.R. designed the studies. J.R.D., A.K., Y.L. and Y.S. conducted the Hi-C experiments; J.R.D., S.S. and F.Y. carried out the data analysis; J.S.L. and M.H. provided insight for analysis; F.Y. built the supporting website; J.R.D. and B.R. prepared the manuscript.

Author Information All Hi-C data described in this study have been deposited in the GEO under accession number GSE35156. We have developed a web-based Java tool to visualize the high-resolution Hi-C data at a genomic region of interest that is available at http://chromosome.sdsc.edu/mouse/hi-c/. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.R. (biren@ucsd.edu).

doi:10.1038/nature11076

Reach and grasp by people with tetraplegia using a neurally controlled robotic arm

Leigh R. Hochberg[1,2,3,4], Daniel Bacher[2]*, Beata Jarosiewicz[1,5]*, Nicolas Y. Masse[5]*, John D. Simeral[1,2,3]*, Joern Vogel[6]*, Sami Haddadin[6], Jie Liu[1,2], Sydney S. Cash[3,4], Patrick van der Smagt[6] & John P. Donoghue[1,2,5]

Paralysis following spinal cord injury, brainstem stroke, amyotrophic lateral sclerosis and other disorders can disconnect the brain from the body, eliminating the ability to perform volitional movements. A neural interface system[1–5] could restore mobility and independence for people with paralysis by translating neuronal activity directly into control signals for assistive devices. We have previously shown that people with long-standing tetraplegia can use a neural interface system to move and click a computer cursor and to control physical devices[6–8]. Able-bodied monkeys have used a neural interface system to control a robotic arm[9], but it is unknown whether people with profound upper extremity paralysis or limb loss could use cortical neuronal ensemble signals to direct useful arm actions. Here we demonstrate the ability of two people with long-standing tetraplegia to use neural interface system-based control of a robotic arm to perform three-dimensional reach and grasp movements. Participants controlled the arm and hand over a broad space without explicit training, using signals decoded from a small, local population of motor cortex (MI) neurons recorded from a 96-channel microelectrode array. One of the study participants, implanted with the sensor 5 years earlier, also used a robotic arm to drink coffee from a bottle. Although robotic reach and grasp actions were not as fast or accurate as those of an able-bodied person, our results demonstrate the feasibility for people with tetraplegia, years after injury to the central nervous system, to recreate useful multidimensional control of complex devices directly from a small sample of neural signals.

The study participants, referred to as S3 and T2 (a 58-year-old woman, and a 66-year-old man, respectively), were each tetraplegic and anarthric as a result of a brainstem stroke. Both were enrolled in the BrainGate2 pilot clinical trial (see Methods). Neural signals were recorded using a 4 mm × 4 mm, 96-channel microelectrode array, which was implanted in the dominant MI hand area (for S3, in November 2005, 5.3 years before the beginning of this study; for T2, in June 2011, 5 months before this study). Participants performed sessions on a near-weekly basis to perform point and click actions of a computer cursor using decoded MI ensemble spiking signals[7]. Across four sessions in her sixth year after implant (trial days 1952–1975), S3 used these neural signals to perform reach and grasp movements of either of two differently purposed right-handed robot arms. The DLR Light-Weight Robot III (German Aerospace Center, Oberpfaffenhofen, Germany; Fig. 1b, left)[10] is designed to be an assistive device that can reproduce complex arm and hand actions. The DEKA Arm System (DEKA Research and Development; Fig. 1b, right) is a prototype advanced upper limb replacement for people with arm amputation[11]. T2 controlled the DEKA prosthetic limb on one session day (day 166). Both robots were operated under continuous user-driven neuronal ensemble control of arm endpoint (hand) velocity in three-dimensional space; a simultaneously decoded neural state executed a hand action. S3 had used the DLR robot on multiple occasions over the previous year for algorithm development and interface testing, but she had no exposure to the DEKA arm before the sessions reported here. T2 participated in three DEKA arm sessions for similar development and testing before the session reported here but had no other experience using the robotic arms.

To decode movement intentions from neural activity, electrical potentials from each of the 96 channels were filtered to reveal extracellular action potentials (that is, ‘unit’ activity). Unit threshold crossings (see Methods) were used to calibrate decoders that generated velocity and hand state commands. Signals for reach were decoded using a Kalman filter[12] to update continuously an estimate of the participant’s intended hand velocity. The Kalman filter was initialized during a single ‘open-loop’ filter calibration block (<4 min) in which the participants were asked to imagine controlling the robotic arm as they watched it undergo a series of regular, pre-programmed movements while the accompanying neural activity was recorded. This open-loop filter was then iteratively updated during four to eight ‘closed-loop’ calibration blocks while the participant actively controlled the robot under visual feedback, with gradually decreasing levels of computer-imposed error attenuation (see Methods). To discriminate an intended hand state, a linear discriminant classifier was built on signals from the same recorded units while the participant imagined squeezing their hand[8]. On average, the decoder calibration procedure lasted ~31 min (ranging from 20 to 48 min, exclusive of time between blocks).

After decoder calibration, we assessed whether each participant could use the robotic arm to reach for and grasp foam ball targets of diameter 6 cm, presented in three-dimensional space one at a time by motorized levers (Fig. 1a–c and Supplementary Fig. 1b). Because hand aperture was not much larger than the target size (only 1.3 times larger for DLR, and 1.8 times larger for DEKA) and hand orientation was not under user control, grasping targets required the participant to manoeuvre the arm within a narrow range of approach angles with the hand open while avoiding the target support rod below. Targets were mounted on flexible supports; brushing them with the robotic arm resulted in target displacements. Together, these factors increased task difficulty beyond simple point-to-point movements and frequently required complex curved paths or corrective actions (Fig. 1d and Supplementary Movies 1–3). Trials were judged successful or unsuccessful by two independent visual inspections of video data (see Methods). A successful ‘touch’ trial occurred when the participant contacted the target with the hand; a successful ‘grasp’ trial occurred when the participant closed the hand while any part of the target or the top of its supporting cone was within the volume enclosed by the hand.

In the three-dimensional reach and grasp task, S3 performed 158 trials across four sessions and T2 performed 45 trials in a single session (Table 1 and Fig. 1e, f). S3 touched the target within the allotted time in 48.8% of the DLR and 69.2% of the DEKA trials, and T2 touched the target within the allotted time in 95.6% of trials (Supplementary

[1]Rehabilitation Research & Development Service, Department of Veterans Affairs, Providence, Rhode Island 02908, USA. [2]School of Engineering and Institute for Brain Science, Brown University, Providence, Rhode Island 02912, USA. [3]Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [4]Harvard Medical School, Boston, Massachusetts 02115, USA. [5]Department of Neuroscience and Institute for Brain Science, Brown University, Providence, Rhode Island 02912, USA. [6]German Aerospace Center, Institute of Robotics and Mechatronics (DLR, Oberpfaffenhofen) 82230, Germany.
*These authors contributed equally to this work.

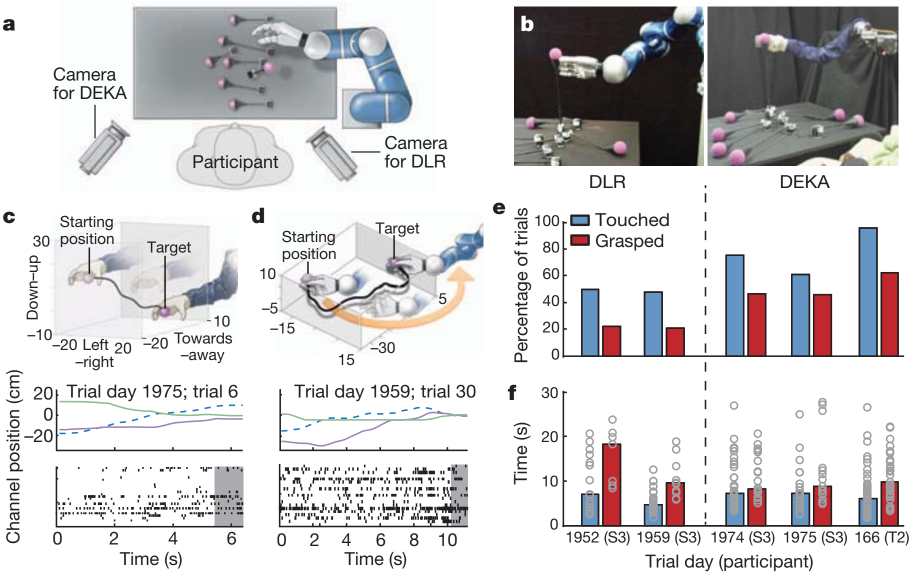


Figure 1 | Experimental setup and performance metrics. **a**, Overhead view of participant's location at the table (grey rectangle) from which the targets (purple spheres) were elevated by a motor. The robotic arm was positioned to the right and slightly in front of the participant (the DLR and DEKA arms were mounted in slightly different locations to maximize the correspondence of their workspaces over the table; for details, see Supplementary Fig. 9). Both video cameras were used for all DLR and DEKA sessions; labels indicate which camera was used for the photographs in **b**. **b**, Photographs of the DLR (left panel) and DEKA (right panel) robots. **c**, Reconstruction of an example trial in which the participant moved the DEKA arm in all three dimensions to reach and grasp a target successfully. The top panel illustrates the trajectory of the hand in three-dimensional space. The middle panel shows the position of the wrist joint for the same trajectory decomposed into each of its three dimensions relative to the participant: the left–right axis (dashed blue line), the towards–away axis (purple line) and the up–down axis (green line). The bottom panel shows the threshold crossing events from all units that contributed to decoding the movement. Each row of tick marks represents the activity of one unit and each tick mark represents a threshold crossing. The grey shaded area shows the first 1 s of the grasp. **d**, An example trajectory from a DLR session in which the participant needed to move the robot hand, which started to the left of the target, around and to the right of the target to approach it with the open part of the hand. The middle and bottom panels are analogous to **c**. **e**, Percentage of trials in which the participant successfully touched the target with the robotic hand (blue bars) and successfully grasped the target (red bars). **f**, Average time required to touch (blue bars) or grasp (red bars) the targets. Each circle shows the acquisition time for one successful trial.

Movies 1–3 and Supplementary Fig. 2). Of the successful touches, S3 grasped the target 43.6% (DLR) and 66.7% (DEKA) of the time, whereas T2 grasped the target 65.1% of the time. Of all trials, S3 grasped the target 21.3% (DLR) and 46.2% (DEKA) of the time, and T2 grasped the target 62.2% of the time. In all sessions from both participants, performance was significantly higher than expected by chance alone (Supplementary Fig. 3). For S3, times to touch were approximately the same for both robotic arms (Fig. 1f, blue bars; median 6.2 ± 5.4 s) and were comparable to times for T2 (6.1 ± 5.5 s). The times for combined reach and grasp were similar for both participants (S3, 9.4 ± 6.2 s; T2, 9.5 ± 5.5 s), although for the first DLR session, times were about twice as long.

To explore the use of neural interface systems for facilitating activities of daily living for people with paralysis, we also assessed how well S3 could control the DLR arm as an assistive device. We asked her to reach for and pick up a bottle of coffee, and then drink from it through a straw and place it back on the table. For this task, we restricted velocity control to the two-dimensional tabletop plane and we used the simultaneously decoded grasp state as a sequentially activated trigger for one of four different hand actions that depended upon the phase of the task and the position of the hand (see Methods). Because the 7.2 cm bottle diameter was 90% of the DLR hand aperture, grasping the bottle required even greater alignment precision than grasping the targets in the three-dimensional task described above. Once triggered by the state switch, robust finger position and grasping of the object was achieved by automated joint impedance control. We familiarized the participant with the task for approximately 14 min (during which we made adjustments to the robot hand grip force, and the participant learned the physical space in which the state decoder and directional commands would be effective in moving the bottle close enough to drink from a straw). After this period, the participant successfully grasped the bottle, brought it to her mouth, drank coffee from it through a straw and replaced the bottle on the table, on four out of six attempts over the next 8.5 min (Fig. 2, Supplementary Fig. 4 and Supplementary Movie 4). The two unsuccessful attempts (numbers 2 and 5 in sequence) were aborted to prevent the arm from pushing the bottle off the table (because the hand aperture was not properly aligned with the bottle). This was the first time since the participant's stroke more than 14 years earlier that she had been able to bring any drinking vessel to her mouth and drink from it solely of her own volition.

The use of neural interface systems to restore functional movement will become practical only if chronically implanted sensors function for many years. It is thus notable that S3's reach and grasp control was achieved using signals from an intracortical array implanted over 5 years earlier. This result, supported by multiple demonstrations of

Table 1 | Summary of neurally controlled robotic arm target-acquisition trials

	Trial day 1952 S3 (DLR)	Trial day 1959 S3 (DLR)	Trial day 1974 S3 (DEKA)	Trial day 1975 S3 (DEKA)	Trial day 166 T2 (DEKA)
Number of trials	32	48	45	33	45
Targets contacted	16 (50.0%)	23 (47.9%)	34 (75.6%)	20 (60.6%)	43 (95.6%)
Grasped	7 (21.9%)	10 (20.8%)	21 (46.7%)	15 (45.5%)	28 (62.2%)
Time to touch (s)	5.4 ± 6.9	5.4 ± 2.3	6.1 ± 4.9	6.8 ± 3.6	5.5 ± 4.7
Time to grasp (s)	18.2 ± 6.4	9.5 ± 4.5	8.2 ± 4.9	8.8 ± 8.0	9.5 ± 5.5
Touched only	9 (28.1%)	13 (27.1%)	13 (28.9%)	5 (15.1%)	15 (33.3%)
Time to touch (s)	7.0 ± 6.2	4.6 ± 3.0	10.7 ± 6.5	9.4 ± 8.0	7.1 ± 6.8

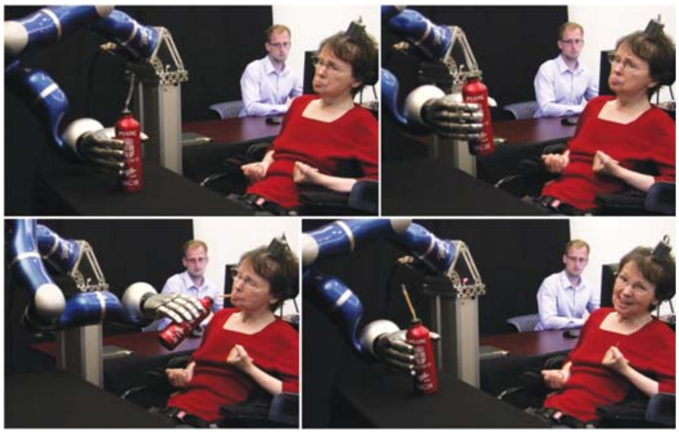


Figure 2 | Participant S3 drinking from a bottle using the DLR robotic arm. Four sequential images from the first successful trial showing participant S3 using the robotic arm to grasp the bottle, bring it towards her mouth, drink coffee from the bottle through a straw (her standard method of drinking) and place the bottle back on the table. The researcher in the background was positioned to monitor the participant and robotic arm. (See Supplementary

successful chronic recording capabilities in animals[13–15], suggests that the goal of creating long-term intracortical interfaces is feasible. At the time of this study, S3 had lower recorded spike amplitudes and fewer channels contributing signals to the filter than during her first years of recording. Nevertheless, the units included in the Kalman filters were sufficiently directionally tuned and modulated to allow neural control of reach and grasp (Fig. 3 and Supplementary Figs 5 and 6). S3 sometimes experiences stereotypic limb flexion. These movements did not appear to contribute in any way to her multidimensional reach and grasp control, and the neural signals used for this control showed waveform shapes and timing characteristics of unit spiking (Fig. 3 and Supplementary Fig. 7). Furthermore, T2 produced no consistent volitional movement during task performance, which further substantiates the intracortical origin of his neural control.

We have shown that two people with no functional arm control due to brainstem stroke used the neuronal ensemble activity generated by intended arm and hand movements to make point-to-point reaches and grasps with a robotic arm across a natural human-arm workspace. Moreover, S3 used these neurally driven commands to perform an everyday task. These findings extend our previous demonstrations of point and click neural control by people with tetraplegia[7,16] and show that neural spiking activity recorded from a small MI intracortical array contains sufficient information to allow people with long-standing tetraplegia to perform even more complex manual skills. This result suggests the feasibility of using cortically driven commands to restore lost arm function for people with paralysis. In addition, we have demonstrated considerably more complex robotic control than previously shown in able-bodied non-human primates[9,17,18]. Both participants operated human-scale arms in a three-dimensional target task that required curved trajectories and precise alignments over a volume that was 1.4–7.7 times greater than has been used by non-human primates. The drinking task, although only two-dimensional + state control, required both careful positioning and correctly timed hand state commands to accomplish the series of actions necessary to retrieve the bottle, drink from it and return it to the table.

Both participants performed these multidimensional actions after long-standing paralysis. For S3, signals were adequate to achieve control 14 years and 11 months after her stroke, showing that MI neuronal ensemble activity remains functionally engaged despite subcortical damage of descending motor pathways. Future clinical research will be needed to establish whether more signals[19–22], signals from additional or other areas[2,23–25], better decoders, explicit participant training or other advances (see Supplementary Materials) will provide more complex, flexible, independent and natural control. In addition to the robotic assistive device shown here, MI signals might also be used by people with paralysis to reanimate paralysed muscles using functional

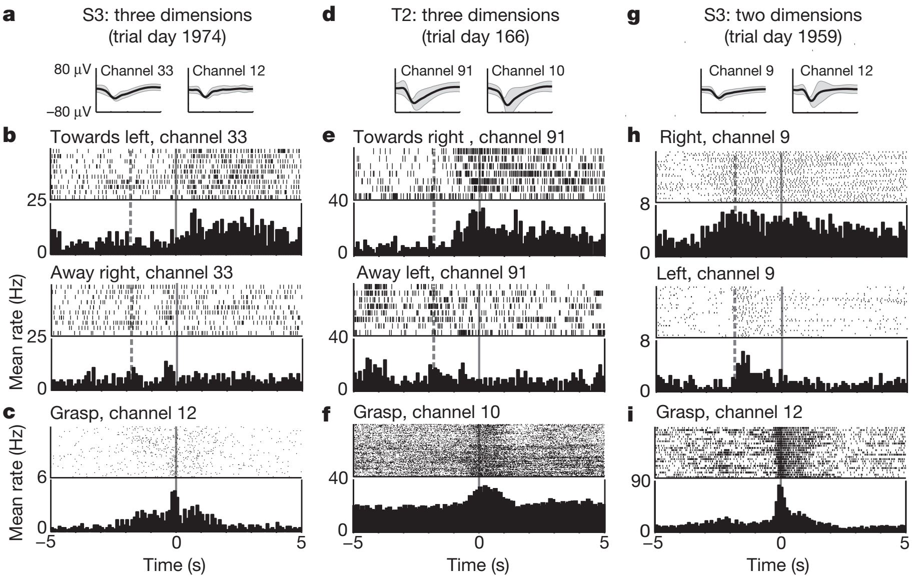


Figure 3 | Examples of neural signals from three sessions and two participants. A three-dimensional reach and grasp session from S3 (**a**–**c**) and T2 (**d**–**f**), and the two-dimensional + grasp drinking session from S3 (**g**–**i**). **a**, **d**, **g**, Average waveforms (black lines) ± two standard deviations (grey shadows) from two units from each session with a large directional modulation of activity. **b**, **e**, **h**, Rasters and histograms of threshold crossings showing directional modulation. Each row of tick marks represents a trial, and each tick mark represents a threshold crossing event. The histogram summarizes the average activity across all trials in that direction. Rasters are displayed for arm movements to and from the pair of opposing targets that most closely aligned with the selected units' preferred directions. Parts **b** and **e** include both closed-loop filter calibration trials and assessment trials; **h** includes only filter calibration trials. Time 0 indicates the start of the trial. The dashed vertical line 1.8 s before the start of the trial identifies the time when the target for the upcoming trial began to rise. Activity occurring before this time corresponded to the end of the previous trial, which often included a grasp, followed by the lowering of the previous target and the computer moving the hand to the next starting position if it was not already there. **c**, **f**, **i**, Rasters and histograms from calibration and assessment trials for units that modulated with intended grasp state. During closed-loop filter calibration trials, the hand automatically closed starting at time 0, cueing the participant to grasp; during assessment trials, the grasp state was decoded at time 0. Expanded data appear in Supplementary Fig. 5.

electrical stimulation[26–28] or by people with limb loss to control prosthetic limbs. Whether MI signals are suitable for people with limb loss to control an advanced prosthetic arm (such as the device shown here) remains to be tested and compared with other control strategies[11,29]. Though further developments might enable people with tetraplegia to achieve rapid, dexterous actions under neural control, at present, for people who have no or limited volitional movement of their own arm, even the basic reach and grasp actions demonstrated here could be substantially liberating, restoring the ability to eat and drink independently.

METHODS SUMMARY

Permission for these studies was granted by the US Food and Drug Administration (Investigational Device Exemption) and the Partners Healthcare/Massachusetts General Hospital Institutional Review Board. Core elements of the investigational BrainGate system have been described previously[6,7].

During each session, participants were seated in a wheelchair with their feet located near or underneath the edge of the table supporting the target placement system. The robotic arm was positioned to the participant's right (Fig. 1a). Raw neural signals for each channel were sampled at 30 kHz and fed through custom Simulink (Mathworks) software in 100 ms bins (S3) or 20 ms bins (T2) to extract threshold crossing rates[2,30]; these threshold crossing rates were used as the neural features for real-time decoding and for filter calibration. Open- and closed-loop filter calibration was performed over several blocks, which were each 3–6 min long and contained 18–24 trials. Targets were presented using a custom, automated target placement platform. On each trial, one of seven servos placed its target (a 6 cm diameter foam ball supported by a spring-loaded wooden dowel rod attached to the servo) in the workspace by lifting it to its task-defined target location (Fig. 1b). Between trials, the previous trial's target was returned to the tabletop while the next target was raised. Owing to variability in the position of the target-placing platform from session to session and changes in the angles of the spring-loaded rods used to hold the targets, visual inspection was used for scoring successful grasp and successful touch trials. Further details on session setup, signal processing, filter calibration, robot systems and target presentations are given in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 September 2011; accepted 26 March 2012.

1. Donoghue, J. P. Bridging the brain to the world: a perspective on neural interface systems. *Neuron* **60,** 511–521 (2008).
2. Gilja, V. *et al.* Challenges and opportunities for next-generation intra-cortically based neural prostheses. *IEEE Trans. Biomed. Eng.* **58,** 1891–1899 (2011).
3. Schwartz, A. B., Cui, X. T., Weber, D. J. & Moran, D. W. Brain-controlled interfaces: movement restoration with neural prosthetics. *Neuron* **52,** 205–220 (2006).
4. Nicolelis, M. A. L. & Lebedev, M. A. Principles of neural ensemble physiology underlying the operation of brain-machine interfaces. *Nature Rev. Neurosci.* **10,** 530–540 (2009).
5. Green, A. M. & Kalaska, J. F. Learning to move machines with the mind. *Trends Neurosci.* **34,** 61–75 (2011).
6. Hochberg, L. R. *et al.* Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature* **442,** 164–171 (2006).
7. Simeral, J. D., Kim, S. P., Black, M. J., Donoghue, J. P. & Hochberg, L. R. Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array. *J. Neural Eng.* **8,** 025027 (2011).
8. Kim, S. P. *et al.* Point-and-click cursor control with an intracortical neural interface system by humans with tetraplegia. *IEEE Trans. Neural Syst. Rehabil. Eng.* **19,** 193–203 (2011).
9. Velliste, M., Perel, S., Spalding, M. C., Whitford, A. S. & Schwartz, A. B. Cortical control of a prosthetic arm for self-feeding. *Nature* **453,** 1098–1101 (2008).
10. Albu-Schaffer, A. *et al.* The DLR lightweight robot: design and control concepts for robots in human environments. *Ind. Rob.* **34,** 376–385 (2007).
11. Resnik, L. Research update: VA study to optimize the DEKA Arm. *J. Rehabil. Res. Dev.* 47**,** ix– x (2010).
12. Wu, W., Gao, Y., Bienenstock, E., Donoghue, J. P. & Black, M. J. Bayesian population decoding of motor cortical activity using a Kalman filter. *Neural Comput.* **18,** 80–118 (2006).
13. Suner, S., Fellows, M. R., Vargas-Irwin, C., Nakata, G. K. & Donoghue, J. P. Reliability of signals from a chronically implanted, silicon-based electrode array in non-human primate primary motor cortex. *IEEE Trans. Neural Syst. Rehabil. Eng.* **13,** 524–541 (2005).
14. Chestek, C. A. *et al.* Long-term stability of neural prosthetic control signals from silicon cortical arrays in rhesus macaque motor cortex. *J. Neural. Eng.* **8,** 045005 (2011).
15. Kruger, J., Caruana, F., Volta, R. D. & Rizzolatti, G. Seven years of recording from monkey cortex with a chronically implanted multiple microelectrode. *Front. Neuroeng.* **3,** 6 (2010).
16. Kim, S. P., Simeral, J. D., Hochberg, L. R., Donoghue, J. P. & Black, M. J. Neural control of computer cursor velocity by decoding motor cortical spiking activity in humans with tetraplegia. *J. Neural Eng.* **5,** 455–476 (2008).
17. Burrow, M., Dugger, J., Humphrey, D. R., Reed, D. J. & Hochberg, L. R. in *Proc. ICORR '97: Int. Conf. Rehabilitation Robotics* 83–86 (Bath Institute of Medical Engineering, 1997).
18. Shin, H. C., Aggarwal, V., Acharya, S., Schieber, M. H. & Thakor, N. V. Neural decoding of finger movements using Skellam-based maximum-likelihood decoding. *IEEE Trans. Biomed. Eng.* **57,** 754–760 (2010).
19. Vargas-Irwin, C. E. *et al.* Decoding complete reach and grasp actions from local primary motor cortex populations. *J. Neurosci.* **30,** 9659–9669 (2010).
20. Mehring, C. *et al.* Inference of hand movements from local field potentials in monkey motor cortex. *Nat. Neurosci.* **6,** 1253–1254 (2003).
21. Stark, E. & Abeles, M. Predicting movement from multiunit activity. *J. Neurosci.* **27,** 8387–8394 (2007).
22. Bansal, A. K., Vargas-Irwin, C. E., Truccolo, W. & Donoghue, J. P. Relationships among low-frequency local field potentials, spiking activity, and three-dimensional reach and grasp kinematics in primary motor and ventral premotor cortices. *J. Neurophysiol.* **105,** 1603–1619 (2011).
23. Musallam, S., Corneil, B. D., Greger, B., Scherberger, H. & Andersen, R. A. Cognitive control signals for neural prosthetics. *Science* **305,** 258–262 (2004).
24. Mulliken, G. H., Musallam, S. & Andersen, R. A. Decoding trajectories from posterior parietal cortex ensembles. *J. Neurosci.* **28,** 12913–12926 (2008).
25. Santhanam, G., Ryu, S. I., Yu, B. M., Afshar, A. & Shenoy, K. V. A high-performance brain–computer interface. *Nature* **442,** 195–198 (2006).
26. Moritz, C. T., Perlmutter, S. I. & Fetz, E. E. Direct control of paralysed muscles by cortical neurons. *Nature* **456,** 639–642 (2008).
27. Pohlmeyer, E. A. *et al.* Toward the restoration of hand use to a paralyzed monkey: brain-controlled functional electrical stimulation of forearm muscles. *PLoS One* **4,** e5924 (2009).
28. Chadwick, E. K. *et al.* Continuous neuronal ensemble control of simulated arm reaching by a human with tetraplegia. *J. Neural Eng.* **8,** 034003 (2011).
29. Kuiken, T. A. *et al.* Targeted reinnervation for enhanced prosthetic arm function in a woman with a proximal amputation: a case study. *Lancet* **369,** 371–380 (2007).
30. Fraser, G. W., Chase, S. M., Whitford, A. & Schwartz, A. B. Control of a brain–computer interface without spike sorting. *J. Neural Eng.* **6,** 055004 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank participants S3 and T2 for their dedication to this research. We thank M. Black for initial guidance in the BrainGate–DLR research. We thank E. Gallivan, E. Berhanu, D. Rosler, L. Barefoot, K. Centrella and B. King for their contributions to this research. We thank G. Friehs and E. Eskandar for their surgical contributions. We thank K. Knoper for assistance with illustrations. We thank D. Van Der Merwe and DEKA Research and Development for their technical support. The contents do not represent the views of the Department of Veterans Affairs or the United States Government. The research was supported by the Rehabilitation Research and Development Service, Office of Research and Development, Department of Veterans Affairs (Merit Review Awards B6453R and A6779I; Career Development Transition Award B6310N). Support was also provided by the National Institutes of Health: NINDS/NICHD (RC1HD063931), NIDCD (R01DC009899), NICHD-NCMRR (N01HD53403 and N01HD10018), NIBIB (R01EB007401), NINDS-Javits (NS25074); a Memorandum of Agreement between the Defense Advanced Research Projects Agency (DARPA) and the Department of Veterans Affairs; the Doris Duke Charitable Foundation; the MGH-Deane Institute for Integrated Research on Atrial Fibrillation and Stroke; Katie Samson Foundation; Craig H. Neilsen Foundation; the European Commission's Seventh Framework Programme through the project The Hand Embodied (grant 248587). The pilot clinical trial into which participant S3 was recruited was sponsored in part by Cyberkinetics Neurotechnology Systems (CKI).

Author Contributions J.P.D. and L.R.H. conceived, planned and directed the BrainGate research and the DEKA sessions. J.P.D., L.R.H. and P.v.d.S. conceived, planned and directed the DLR robot control sessions. J.P.D. and P.v.d.S. are co-senior authors. D.B., B.J., N.Y.M., J.D.S. and J.V. contributed equally to this work and are listed alphabetically. J.D.S., J.V. and D.B. developed the BrainGate–DLR interface. D.B., J.D.S. and J.L. developed the BrainGate–DEKA interface. D.B. and J.V. created the three-dimensional motorized target placement system. B.J., N.Y.M. and D.B. designed the behavioural task, the neural signal processing approach, the filter building approach and the performance metrics. B.J., N.Y.M. and D.B. performed data analysis, further guided by L.R.H., J.D.S. and J.P.D. N.Y.M., L.R.H. and J.P.D. drafted the manuscript, which was further edited by all authors. D.B. and J.D.S. engineered the BrainGate neural interface system/assistive technology system. J.V. and S.H. developed the reactive planner for the Light-Weight Robot III (LWR). S.H. developed the internal control framework of the Light-Weight Robot III. The internal control framework of the DEKA arm was developed by DEKA. L.R.H. is principal investigator of the pilot clinical trial. S.S.C. is clinical co-investigator of the pilot clinical trial and assisted in the clinical oversight of these participants.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.P.D. (john_donoghue@brown.edu) or L.R.H. (leigh@brown.edu).

METHODS

Permission for these studies was granted by the US Food and Drug Administration (Investigational Device Exemption) and the Partners Healthcare/Massachusetts General Hospital Institutional Review Board. The two participants in this study, S3 and T2, were enrolled in a pilot clinical trial of the BrainGate Neural Interface System (additional information about the clinical trial is available at http://www.clinicaltrials.gov/ct2/show/NCT00912041).

At the time of this study, S3 was a 58-year-old woman with tetraplegia caused by brainstem stroke that occurred nearly 15 years earlier. As previously reported[7,31], she is unable to speak (anarthria) and has no functional use of her limbs. She has occasional bilateral or asymmetric flexor spasm movements of the arms that are intermittently initiated by any imagined or actual attempt to move. S3's sensory pathways remain intact. She also retains some head movement and facial expression, has intact eye movement and breathes spontaneously. On 30 November 2005, a 96-channel intracortical silicon microelectrode array (1.5 mm electrode length, produced by Cyberkinetics Neurotechnology Systems, and now by its successor, Blackrock Microsystems) was implanted in the arm area of motor cortex as previously described[6,7]. One month later, S3 began regularly participating in one or two research sessions per week during which neural signals were recorded and tasks were performed towards the development, assessment and improvement of the neural interface system. The data reported here are from S3's trial days 1952–1975, more than 5 years after implant of the array. Participant S3 provided permission for photographs, videos and portions of her protected health information to be published for scientific and educational purposes.

The second study participant, T2, was, at the time of this study, a 66-year-old ambidextrous man with tetraplegia and anarthria as a result of a brainstem stroke that occurred in 2006, five and a half years before the collection of the data presented in this report. He has a tracheostomy and percutaneous gastrostomy tube; he receives supportive mechanical ventilation at night but breathes without assistance during the day, and receives all nutrition by percutaneous gastrostomy. He has a left abducens palsy with intermittent diplopia. He can rotate his head slowly over a limited range of motion. With the exception of unreliable and trace right wrist and index finger extension (but not flexion), he is without voluntary movement at and below C5. Occasional coughing results in involuntary hip flexion, and intermittent, rhythmic chewing movements occur without alteration in consciousness. Participant T2 also had a 96-channel Blackrock array with 1.5 mm electrodes implanted into the dominant arm–hand area of motor cortex; the array was placed 5 months before the session reported here.

Setup. During each session, the participant was seated in their wheelchair with their feet located underneath the edge of the table supporting the target placement system. The robot arm was positioned to the participant's right (Fig. 1a). A technician used aseptic technique to connect the 96-channel recording cable to the percutaneous pedestal and then viewed neural signal waveforms using commercial software (Cerebus Central, Blackrock Microsystems). The waveforms were used to identify channels that were not recording signals and/or were contaminated with noise; for S3, those channels were manually excluded and remained off for the remainder of the recording session.

Robot systems. We used two robot systems with multi-joint arms and hands during this study. The first was the DLR Light-Weight Robot III[10,32], with the DLR Five-Finger Hand[33], developed at the German Aerospace Center (DLR). The arm weighs 14 kg and has seven degrees of freedom (DoF). The hand has 15 active DoF which were combined into a single DoF (hand open/close) to execute a grasp for these experimental sessions. Torque sensors were embedded in each joint of the arm and hand, allowing the system to operate under impedance control, and enabling it to handle collision safely, which is desirable for human–robot interactions[34]. The hand orientation was fixed in Cartesian space. The second robotic system was the DEKA Generation 2 prosthetic arm system, which weighs 3.64 kg and has six DoF in the arm (shoulder abduction, shoulder flexion, humeral rotation and elbow flexion, wrist flexion, wrist rotation), and four DoF in the hand (also combined into a single DoF to execute a grasp for these experimental sessions). The DEKA hand orientation was kept fixed in joint space; therefore, it could change in the Cartesian space depending upon the posture of other joints derived from the inverse kinematics.

Both robotic arms were controlled in endpoint velocity space while a parallel state switch, also under neural control from the same cortical ensemble, controlled grasp. Virtual boundaries were placed in the workspace as part of the control software to avoid collisions with the tabletop, support stand and participant. Of the 158 trials performed by S3, 80 were performed during the first two sessions using the DLR arm and 78 during the two sessions using the DEKA arm.

Target presentation. Targets were defined using a custom, automated servo-based robotic platform. On each trial, one of the seven servos placed its target (a 6 cm diameter foam ball attached to the servo by a spring-loaded wooden dowel rod) in the workspace by lifting it to its task-defined target location. Between trials, the previous target was returned to the table while the next target was raised to its position. The trials alternated between the lower right 'home' target and one of the other six targets. The targets circumscribed an area of 30 cm from left to right, 52 cm in depth and 23 cm vertically (see Supplementary Figs 1 and 9).

Owing to variability in the position of the target-placing platform from session to session and changes in the angles of the spring-loaded rods used to hold the targets, estimates of true target locations in physical space relative to the software-defined targets were not exact. This target placement error had no impact on the three-dimensional reach and grasp task because the goal of the task was to grab the physical target regardless of its exact location. However, for this reason, it was not possible to use an automated method for scoring touches and grasps. Instead, scoring was performed by visual inspection of the videos: for S3, by a group of three investigators (N.Y.M., D.B. and B.J.) and independently by a fourth investigator (L.R.H.); for T2, independently by four investigators (J.D.S., D.B., and B.J. and L.R.H.). Of 203 trials, there was initial concordance in scoring in 190 of them. The remaining 13 were re-reviewed using a second video taken from a different camera angle, and either a unanimous decision was reached ($n = 10$) or when there was any unresolved discordance in voting, the more conservative score was assigned ($n = 3$).

Signal acquisition. Raw neural signals for each channel were sampled at 30 kHz and fed through custom Simulink (Mathworks) software in 100 ms bins (for participant S3) or 20 ms bins (for participant T2). For participant T2, coincident noise in the raw signal was reduced using common-average referencing: from the 50 channels with the lowest impedance, we selected the 20 with the lowest firing rates. The mean signal from these 20 channels was subtracted from all 96 channels.

To extract threshold crossing rates[2,30], signals in each bin were then filtered with a fourth-order Butterworth filter with corners at 250 and 5,000 Hz, temporally reversed and filtered again. Neural signals were buffered for 4 ms before filtering to avoid edge effects. This symmetric (non-causal) filter is better matched to the shape of a typical action potential[35], and using this method led to better extraction of low-amplitude action potentials from background noise and higher directional modulation indices than would be obtained using a causal filter. Threshold crossings were counted as follows. For computational efficiency, signals were divided into 2.5 ms (for S3) or 0.33 ms (for T2) sub-bins, and in each sub-bin, the minimum value was calculated and compared with a threshold. For S3, this threshold was set at -4.5 times the filtered signal's root mean square value in the previous block. For T2, this threshold was set at -5.5 times the root mean square of the distribution of minimum values collected from each sub-bin. (Offline analysis showed that these two methods produced similar threshold values relative to noise amplitude.) To prevent large spike amplitudes from inflating the root mean square estimate for both S3 and T2, signal values were capped between 40 μV and -40 μV before calculating this threshold for each channel. The number of minima that exceeded the channel's threshold was then counted in each bin, and these threshold crossing rates were used as the neural features for real-time decoding and for closed-loop filter calibration.

Filter calibration. Filter calibration was performed at the beginning of each session using data acquired over several 'blocks' of 18–24 trials (each block lasting approximately 3–6 min). The process began with one open-loop filter initialization block, in which the participants were instructed to imagine that they were controlling the movements of the robot arm as it performed pre-programmed movements along the cardinal axes. The trial sequence was a centre–out–back pattern. Each block began with the endpoint of the robot arm at the 'home' target in the middle of the workspace. The hand would then move to a randomly selected target (distributed equidistant from the home target on the cardinal axes), pause there for 2 s, then move back to the home target. This pattern was repeated two or three times for each target. To initialize the Kalman filter[12,36], a tuning function was estimated for each unit by regressing its threshold crossing rates against instantaneous target directions (see below). For participant T2, a 0.3 s exponential smoothing filter was applied to the threshold crossing rates before filter calibration.

Open-loop filter initialization was followed by several blocks of closed-loop filter calibration (adapted to the Kalman filter from refs 37 and 38), in which the participant actively controlled the robot to acquire targets, in a similar home–out–back pattern, but with the home target at the right of the workspace (Supplementary Fig. 1). In each closed-loop filter calibration block, the error in the participant's decoded trajectories was attenuated by scaling down decoded movement commands orthogonal to the instantaneous target direction by a fixed percentage, similar to the technique used by Velliste *et al.*[9]. The amount of error attenuation was decreased across filter calibration blocks until it was zero, giving the participant full three-dimensional control of the robot.

During each closed-loop filter calibration block, the participant's intended movement direction at each moment was inferred to be from the current endpoint of the robot hand towards the centre of the target. Time bins from 0.2 to 3.2 s after the trial start were used to calculate tuning functions and the baseline rates (see below) by regressing threshold crossing rates from each bin against the

corresponding unit vector pointing in the intended movement direction; using this time period was meant to isolate the initial portion of each trial, during which the participant's intended movement direction was less likely to be influenced by error correction. Times when the endpoint was within 6 cm of the target were also excluded, because angular error in the estimation of the intended direction is magnified as the endpoint gets closer to the target.

The state decoder used to control the grasping action of the robot hand was also calibrated during the same open- and closed-loop blocks. During open-loop blocks, after each trial ending at the home target, the robot hand would close for 2 s. During this time, the participant was instructed to imagine that they were closing their own hand. State decoder calibration was similar during closed-loop calibration blocks: after each home target trial, the hand moved to the home target if the participant had not already moved it there, and an auditory cue instructed the participant to imagine closing their own hand. In closed-loop grasp calibration blocks using the DLR arm, the robot hand would only close if the state decoder successfully detected a grasp intention from the participant's neural activity. In closed-loop calibration blocks using the DEKA arm, the hand always closed during grasp calibration irrespective of the decoded grasp state.

Sequential activation of DLR robot hand actions during the drinking task. In the drinking task, when participant S3 activated a grasp state, one of four different hand/arm actions were activated, depending upon the phase of the task and the position of the hand: (1) close the hand around the bottle and raise it off the table; (2) stop arm movement and pronate the wrist to orient the bottle towards the participant; (3) supinate the wrist back to its original position and re-enable arm movement; or (4) lower the bottle to the table and withdraw the hand.

Tracking baseline firing rates. Endpoint velocity and grasp state were decoded based on the deviation of each unit's neural activity from its baseline rate; thus, errors in estimating the baseline rate itself may create a bias in the decoded velocity or grasp state. To reduce such biases despite potential drifts in baseline rates over time, the baseline rates were re-estimated after every block using the previous block's data.

During filter calibration, in which the participant was instructed to move the endpoint of the hand directly towards the target, we determined the baseline rate of a channel by modelling neural activity as a linear function of the intended movement direction plus the baseline rate. Specifically, the following equation was fitted: $z = \text{baseline} + Hd$, where z is the threshold crossing rate, H is the channel's preferred direction and d is the intended movement direction. As described above for the filter calibration, only data during the initial portion of the trial, from 0.2 to 3.2 s after trial start, were used to fit the model. Only the last block's data were used to estimate each unit's baseline rate for use during decoding in the following block (unless the last block was aborted for a technical reason, in which case the baseline rates were taken from the last full block).

This method for baseline rate tracking was not used for S3's drinking demonstration or for the blocks in which the participant was instructed to reach and grasp the targets because it could no longer be assumed that the participant was intending to move the endpoint of the hand directly towards the target (Fig. 1d). For these blocks, the mean threshold crossing rate of each unit across the entire block was used as a proxy for its baseline rate. Mean rates did not differ substantially from baseline rates calculated from the same block (data not shown).

Hand velocity and grasp filters. During closed-loop blocks, the endpoint velocity of the robot arm and the state of the hand were controlled in parallel by decoded neural activity, and were updated every 100 ms for S3, and every 20 ms for T2. The desired endpoint velocity was decoded using a Kalman filter[7,8,12,36]. The Kalman filter requires four sets of parameters, two of which were calculated based on the mean-subtracted (and for T2, smoothed with a 0.3 s exponential filter) threshold crossing rate, $\bar{z}$, and the intended direction, d, whereas the other two parameters were hard coded. The first parameter was the directional tuning, H, calculated as $H = \bar{z}d^T(dd^T)^{-1}$. The second parameter, Q, was the error covariance matrix in linearly reconstructing the neural activity, $Q = (\bar{z} - Hd)(\bar{z} - Hd)^T$. The two hard-coded parameters were the state transition matrix A, which predicts the intended direction given the previous estimate $d(t) = Ad(t-1)$, and the error in this model,

$$W = \frac{1}{N}\sum_{t=1}^{N}(d(t) - Ad(t-1))(d(t) - Ad(t-1))^T.$$

These values were set to $A = 0.965I$ for both S3 and T2, and $W = 0.03I$ for S3 and $W = 0.012I$ for T2, where I is the identity matrix (W was set to a lower value for T2 to achieve a similar endpoint 'inertia' as for S3 despite the smaller bin size used for T2). From past experience, it was found that fitting these two parameters from the perfectly smoothed open-loop kinematics data produced too much inertia in the commanded movement to control the robot arm properly, though this may have been a function of the relative paucity of signals rather than a suboptimal component of the decoding algorithm.

To select channels to be included in the filter, we first defined a 'modulation index' as the magnitude of a unit's modelled preferred direction vector (that is, the amplitude of its cosine fit from baseline to peak rate), in hertz. When unit vectors are used for the intended movement direction in the filter calibration regression, this is equivalent to $\|H_i\|$, where H_i is the row of the tuning model matrix H that corresponds to channel i. We further defined a 'normalized modulation index' as the modulation index normalized by the standard deviation of the residuals of the unit's cosine fit. Thus, a unit with no directional tuning would have normalized modulation index of 0, a unit whose directional modulation is equal to the standard deviation of its residuals would have a normalized modulation index of 1, and a unit whose directional modulation is larger than the standard deviation of its residuals would have a normalized modulation index greater than 1. We included all channels with baseline rates below 100 Hz and with normalized modulation indices above 0.1 for S3 and 0.05 for T2. For T2, we included a maximum of 50 channels; channels with the lowest normalized modulation indices were excluded if this limit was exceeded. Across the six sessions, the number of channels included in the Kalman filter ranged from 13 to 50 (see Supplementary Table 1 and Supplementary Fig. 8).

The state decoder used for hand grasp was built using similar methods, as previously described[8]. Briefly, threshold crossings were summed over the previous 300 ms, and linear discriminant analysis was used to separate threshold crossing counts arising when the participant was intending to close the hand from times that they were imagining moving the arm. For the state decoder, we used all channels that were not turned off at the start of the session (see Setup in Methods) and whose baseline threshold crossing rates, calculated from the previous block, were between 0.5 and 100 Hz. Additionally for T2, we only included channels if the difference in mean rates during grasp versus move states divided by the firing rate standard deviation (the d' score) was above 0.05. As for the Kalman filter, we included a maximum of 50 channels in the state decoder for T2; channels with the lowest d' scores were excluded if this limit was exceeded. Across the six sessions, the number of channels included in the state decoder ranged from 16 to 50 (see Supplementary Table 1). Immediately after a grasp was decoded, the Kalman prior was reset to zero. For both robot systems, at the end of a trial, velocity commands were suspended and the arm was repositioned under computer control to the software-expected position of the current target, to prepare the arm to enable the collection of metrics for the next three-dimensional point-to-point reach. Additionally, during the DEKA sessions, three-dimensional velocity commands were suspended during grasps (which lasted 2 s).

Bias correction. For T2, a bias correction method was implemented to reduce biases in the decoded velocity caused by within-block non-stationarities in the neural signals. At each moment, the velocity bias was estimated by computing an exponentially weighted running mean (with a 30 s time constant) of all decoded velocities whose speeds exceeded a predefined threshold. The threshold was set to the 66th centile of the decoded speeds estimated during the most recent filter calibration, which was empirically found to be high enough to include movements caused by biases as well as 'true' high-velocity movements, but importantly, to exclude low-velocity movements generated in an effort to counteract any existing biases. This exponentially weighted running mean was subtracted from the decoded velocity signals to generate a bias-corrected velocity that commanded the endpoint of the DEKA arm.

31. Kim, S. P. *et al.* Multi-state decoding of point-and-click control signals from motor cortical activity in a human with tetraplegia. *3rd Int. IEEE/EMBS Conf. Neural Eng.* 486–489 (2007).
32. Albu-Schaffer, A. *et al.* Soft robotics: from torque feedback controlled light-weight robots to intrinsically compliant systems. *Robot. Automat. Mag.* **15,** 20–30 (2008).
33. Liu, H. *et al.* Multisensory five-finger dexterous hand: The DLR/HIT Hand II. *IEEE/RSJ Int. Conf. Intell. Robots Systems* 3692–3697 (2008).
34. Haddadin, S., Albu-Schaeffer, A. & Hirzinger, G. Requirements for safe robots: measurements, analysis and new insights. *Int. J. Robot. Res.* **28,** 1507–1527 (2009).
35. Quian Quiroga, R. What is the real shape of extracellular spikes? *J. Neurosci. Methods* **177,** 194–198 (2009).
36. Malik, W. Q., Truccolo, W., Brown, E. N. & Hochberg, L. R. Efficient decoding with steady-state Kalman filter in neural interface systems. *IEEE Trans. Neural Syst. Rehabil. Eng.* **19,** 25–34 (2011).
37. Taylor, D. M., Tillery, S. I. & Schwartz, A. B. Direct cortical control of 3D neuroprosthetic devices. *Science* **296,** 1829–1832 (2002).
38. Jarosiewicz, B. *et al.* Functional network reorganization during learning in a brain-computer interface paradigm. *Proc. Natl Acad. Sci. USA* **105,** 19486–19491 (2008).

doi:10.1038/nature11049

Spatial partitioning of the regulatory landscape of the X-inactivation centre

Elphège P. Nora[1,2,3], Bryan R. Lajoie[4]*, Edda G. Schulz[1,2,3]*, Luca Giorgetti[1,2,3]*, Ikuhiro Okamoto[1,2,3], Nicolas Servant[1,5,6], Tristan Piolot[1,2,3], Nynke L. van Berkum[4], Johannes Meisig[7], John Sedat[8], Joost Gribnau[9], Emmanuel Barillot[1,5,6], Nils Blüthgen[7], Job Dekker[4] & Edith Heard[1,2,3]

In eukaryotes transcriptional regulation often involves multiple long-range elements and is influenced by the genomic environment[1]. A prime example of this concerns the mouse X-inactivation centre (*Xic*), which orchestrates the initiation of X-chromosome inactivation (XCI) by controlling the expression of the non-protein-coding *Xist* transcript. The extent of *Xic* sequences required for the proper regulation of *Xist* remains unknown. Here we use chromosome conformation capture carbon-copy (5C)[2] and super-resolution microscopy to analyse the spatial organization of a 4.5-megabases (Mb) region including *Xist*. We discover a series of discrete 200-kilobase to 1 Mb topologically associating domains (TADs), present both before and after cell differentiation and on the active and inactive X. TADs align with, but do not rely on, several domain-wide features of the epigenome, such as H3K27me3 or H3K9me2 blocks and lamina-associated domains. TADs also align with coordinately regulated gene clusters. Disruption of a TAD boundary causes ectopic chromosomal contacts and long-range transcriptional misregulation. The *Xist*/*Tsix* sense/antisense unit illustrates how TADs enable the spatial segregation of oppositely regulated chromosomal neighbourhoods, with the respective promoters of *Xist* and *Tsix* lying in adjacent TADs, each containing their known positive regulators. We identify a novel distal regulatory region of *Tsix* within its TAD, which produces a long intervening RNA, *Linx*. In addition to uncovering a new principle of *cis*-regulatory architecture of mammalian chromosomes, our study sets the stage for the full genetic dissection of the X-inactivation centre.

The X-inactivation centre was originally defined by deletions and translocations as a region spanning several megabases[3,4], and contains several elements known to affect *Xist* activity, including its repressive antisense transcript *Tsix* and its regulators *Xite*, *DXPas34* and *Tsx*[5,6]. However, additional control elements must exist, as single-copy transgenes encompassing *Xist* and up to 460 kb of flanking sequences are unable to recapitulate proper *Xist* regulation[7]. To characterize the *cis*-regulatory landscape of the *Xic* in an unbiased approach, we performed 5C[2] across a 4.5-Mb region containing *Xist*. We designed 5C-Forward and 5C-Reverse oligonucleotides following an alternating scheme[2], thereby simultaneously interrogating nearly 250,000 possible chromosomal contacts in parallel, with a mean resolution of 10–20 kb (Fig. 1a; see Supplementary Methods). Analysis of undifferentiated mouse embryonic stem cells (ESCs) revealed that long-range (>50 kb) contacts preferentially occur within a series of discrete genomic blocks, each covering 0.2–1 Mb (Fig. 1b). These blocks differ from the higher-order organization recently observed by Hi-C[8], corresponding to much larger domains of open or closed chromatin, that come together in the nucleus to form A and B types of compartments[8]. Instead, our 5C analysis shows self-associating chromosomal domains occurring at the sub-megabase scale. The size and location of these domains is identical in male and female mouse ESCs (Supplementary Fig. 1) and in different mouse ESC lines (Supplementary Fig. 2 and Supplementary Data 1).

To examine this organization with an alternative approach, we performed three-dimensional DNA fluorescent *in situ* hybridization (FISH) in male mouse ESCs. Nuclear distances were found to be significantly shorter between probes lying in the same 5C domain than in different domains (Fig. 1c, d), and a strong correlation was found between three-dimensional distances and 5C counts (Supplementary Fig. 3a, b). Furthermore, using pools of tiled bacterial artificial chromosome (BAC) probes spanning up to 1 Mb and structured illumination microscopy, we found that large DNA segments belonging to the same 5C domain colocalize to a greater extent than DNA segments located in adjacent domains (Fig. 1e), and this throughout the cell cycle (Supplementary Fig. 3c, d). Based on 5C and FISH data, we conclude that chromatin folding at the sub-megabase scale is not random, and partitions this chromosomal region into a succession of topologically associating domains (TADs).

We next investigated what might drive chromatin folding in TADs. We first noticed a striking alignment between TADs and the large blocks of H3K27me3 and H3K9me2 (ref. 9) that are known to exist throughout the mammalian genomes[10–13] (for example, TAD E, Fig. 2 and Supplementary Fig. 4). We therefore examined 5C profiles of $G9a^{-/-}$ (also known as *Ehmt2*) mouse ESCs, which lack H3K9me2, notably at the *Xic*[14], and $Eed^{-/-}$ mouse ESCs, which lack H3K27me3 (ref. 15). No obvious change in overall chromatin conformation was observed, and TADs were not affected either in size or position in these mutants (Fig. 2 and Supplementary Fig. 4b). Thus TAD formation is not due to domain-wide H3K27me3 or H3K9me2 enrichment. Instead, such segmental chromatin blocks might actually be delimited by the spatial partitioning of chromosomes into TADs.

We then addressed whether folding in TADs is driven by discrete boundary elements at their borders. 5C was performed in a mouse ESC line carrying a 58-kb deletion (ΔXTX[16]), encompassing the boundary between the *Xist* and *Tsix* TADs (D and E; Fig. 2b). We observed ectopic contacts between sequences in TADs D and E and an altered organization of TAD E. Boundary elements can thus mediate the spatial segregation of neighbouring chromosomal segments. Within the TAD D–E boundary, a CTCF-binding site was recently implicated in insulating *Tsix* from remote regulatory influences[17]. However, alignment of CTCF- and cohesin-binding sites in mouse ESCs[18] with our 5C data showed that, although these factors are present at most TAD boundaries (Supplementary Fig. 4), they are also frequently present within TADs, excluding them as the sole determinants of TAD

[1]Institut Curie, 26 rue d'Ulm, Paris F-75248, France. [2]CNRS UMR3215, Paris F-75248, France. [3]INSERM U934, Paris F-75248, France. [4]Programs in Systems Biology and Gene Function and Expression, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts 01605-0103, USA. [5]INSERM U900, Paris, F-75248 France. [6]Mines ParisTech, Fontainebleau, F-77300 France. [7]Institute of Pathology, Charité–Universitätsmedizin, 10117 Berlin, and Institute of Theoretical Biology Humboldt Universität, 10115 Berlin, Germany. [8]Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, California 94158-2517, USA. [9]Department of Reproduction and Development, Erasmus MC, University Medical Center, 3000 CA Rotterdam, The Netherlands.
*These authors contributed equally to this work.

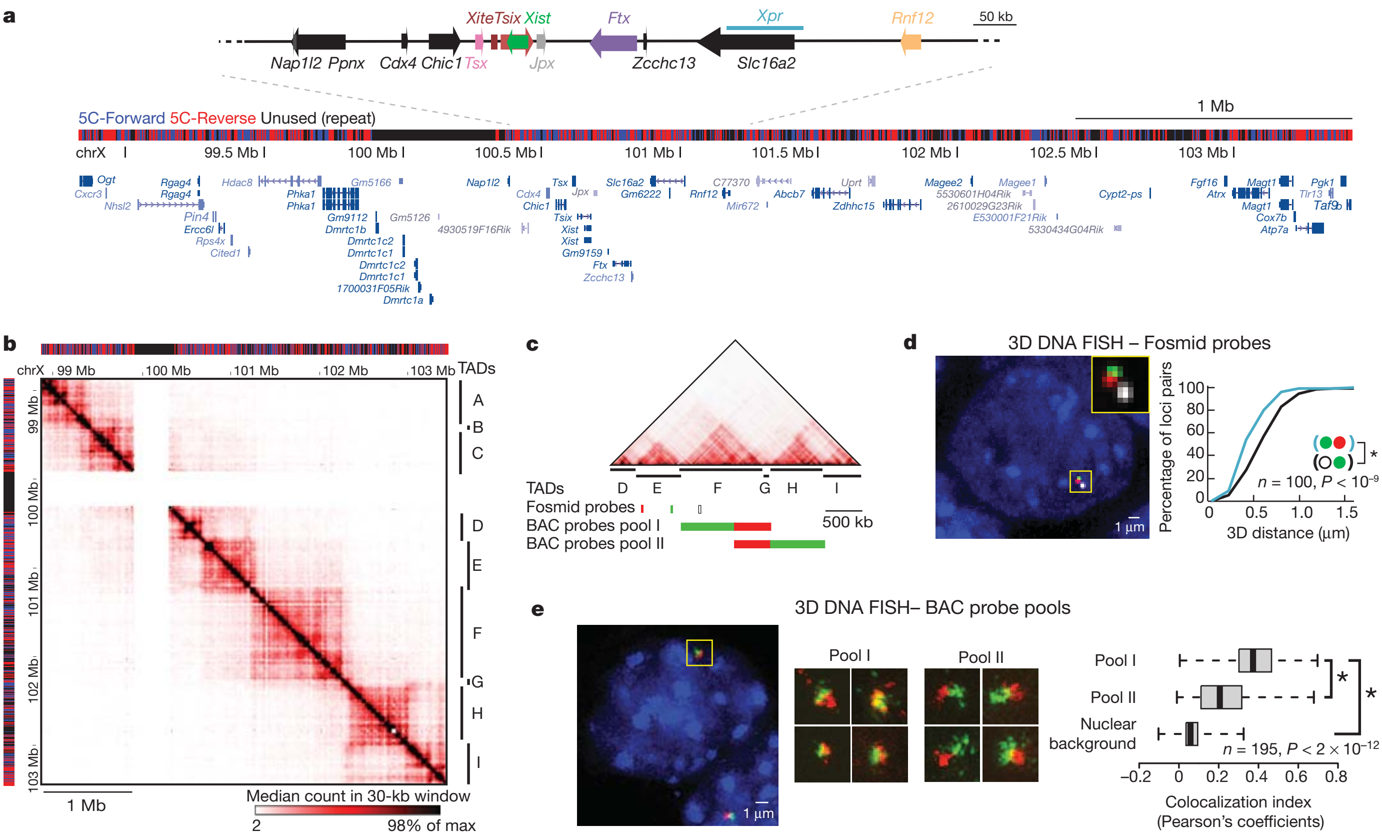


Figure 1 | Chromosome partitioning into topologically associating domains (TADs). **a**, Distribution of 5C-Forward and 5C-Reverse HindIII restriction fragments across the 4.5 Mb analysed showing positions of RefSeq genes and known XCI regulatory loci. **b**, 5C data sets from XY undifferentiated mouse ESCs (E14), displaying median counts in 30-kb windows every 6 kb. Chromosomal contacts are organized into discrete genomic blocks (TADs A–I). A region containing segmental duplications excluded from the 5C analysis is masked (white). **c**, Positions of DNA FISH probes. **d**, Interphase nuclear distances are smaller for probes in the same 5C domain. **e**, Structured illumination microscopy reveals that colocalization of neighbouring sequences is greater when they belong to the same 5C domain. Boxplots show the distribution of Pearson's correlation coefficient between red and green channels, with whiskers and boxes encompassing all and 50% of values, respectively; central bars denote the median correlation coefficient. Statistical significance was assessed using Wilcoxon's rank sum test.

positioning. Furthermore, the fact that the two neighbouring domains do not merge completely in ΔXTX cells (Fig. 2b) implies that additional elements, within TADs, can act as relays when a main boundary is removed. The factors underlying an element's capacity to act as a canonical or shadow boundary remain to be investigated.

Next we asked whether TAD organization changes during differentiation or XCI. Both male neuronal progenitors cells (NPCs) and male primary mouse embryonic fibroblasts (MEFs) show similar organization to mouse ESCs, with no obvious change in TAD positioning. However, consistent differences in the internal contacts within TADs were observed (Fig. 3a, Supplementary Figs 2 and 5). Noticeably, some TADs were found to become lamina-associated domains[19] (LADs) at certain developmental stages (Fig. 3b). Thus chromosome segmentation into TADs reveals a modular framework where changes in chromatin structure or nuclear positioning can occur in a domain-wide fashion during development.

We then assessed TAD organization on the inactive X, by combining *Xist* RNA FISH, to identify the inactive X, and super-resolution DNA FISH using BAC probe pools on female MEFs. We found that colocalization indices on the inactive X were still higher for sequences belonging to the same TAD than for neighbouring TADs (Supplementary Fig. 6a). However, the difference was significantly lower for the inactive X than for the active X. Deconvolution of the respective contributions of the active X and inactive X in 5C data from female MEFs (see Supplementary Methods and Supplementary Fig. 6) similarly revealed that global organization in TADs remains on the inactive X, albeit in a much attenuated form, but that specific long-range contacts within TADs are lost. This, together with a recent report focused on longer-range interactions[20], suggests that the inactive X has a more random chromosomal organization than its active homologue, even below the megabase scale.

We next investigated how TAD organization relates to gene expression dynamics during early differentiation. A transcriptome analysis, consisting of microarray measurements at 17 time points over the first 84 h of female mouse ESC differentiation was performed (Fig. 4a). During this time window, most genes in the 5C region were either up- or downregulated. Statistical analysis demonstrated that expression profiles of genes with promoters located within the same TAD are correlated (Fig. 4b). This correlation (median correlation coefficient cc of 0.40) is significantly higher than for genes in different domains (cc of 0.03, $P < 10^{-9}$) or for genes across the X chromosome in randomly selected, TAD-size regions (cc of 0.09, $P < 10^{-7}$). The observed correlations within TADs seem not to depend on distance between genes, and are thus distinct from previously described correlations between neighbouring genes[21] that decay on a length scale of approximately 100 kb (Supplementary Fig. 7). Our findings indicate that physical clustering within TADs may be used to coordinate gene expression patterns during development. Furthermore, deletion of the boundary between *Xist* and *Tsix* in ΔXTX cells was accompanied by long-range transcriptional misregulation (Supplementary Fig. 8), underlining the role that chromosome partitioning into TADs can play in long-range transcriptional control.

A more detailed analysis of each domain (Supplementary Fig. 7) revealed that co-expression is particularly pronounced in TADs D, E

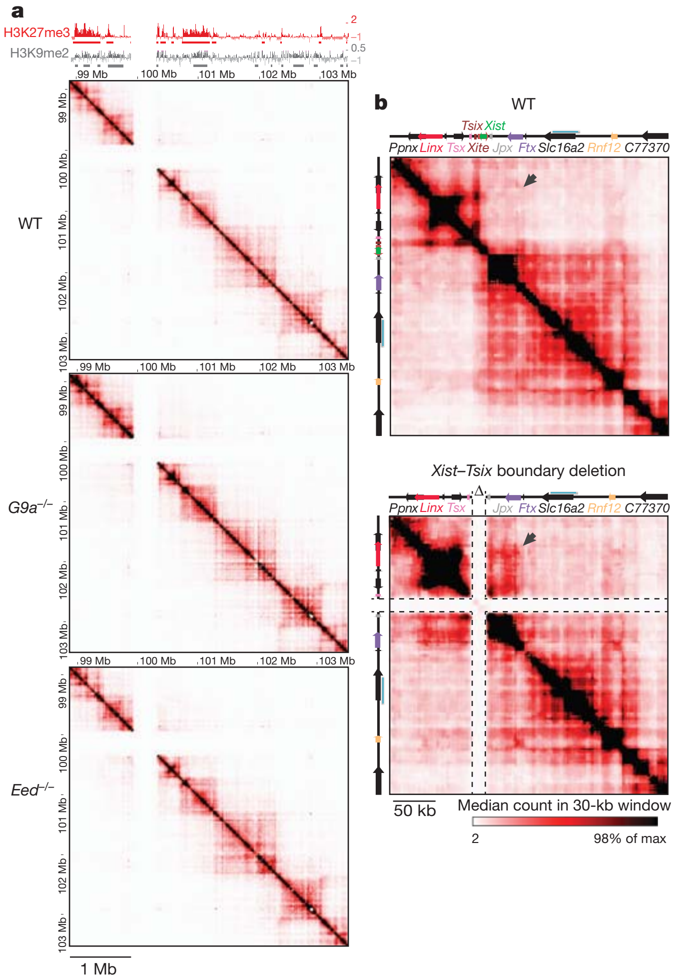


Figure 2 | Determinants of topologically associating domains. **a**, Blocks of contiguous enrichment in H3K27me3 or H3K9me2 (ref. 11) align with the position of TADs (chromatin immunoprecipitation on chip from ref. 9) in wild-type cells (TT2), but TADs are largely unaffected in the absence of H3K9me2 in male *G9a*$^{-/-}$ cells or H3K27me3 in male *Eed*$^{-/-}$ cells. **b**, Deletion of a boundary at *Xist/Tsix* disrupts folding pattern of the two neighbouring TADs.

and F (Fig. 4b, c). Although correlations are strongest within TADs, there is some correlation between TADs showing the same trend, such as TADs D and F, which are both downregulated during differentiation. Only TAD E, which contains *Xist* and all of its known positive regulators *Jpx*, *Ftx*, *Xpr/Xpct* and *Rnf12*[5] (*Jpx*, *Ftx*, *Xpct* and *Rnf12* are also known as *Enox*, *B230206F22Rik*, *Slc16a2* and *Rlim*, respectively) is anti-correlated with most other genes in the 4.5 Mb region, being upregulated during differentiation (Supplementary Fig. 7). The fact that these coordinately upregulated loci are located in the same TAD suggests that they are integrated into a similar *cis*-regulatory network, potentially sharing common *cis*-regulatory elements. We therefore predict that TAD E (~550 kb) represents the minimum 5′ regulatory region required for accurate *Xist* expression, explaining why even the largest transgenes tested so far (covering 150 kb 5′ to *Xist*, Fig. 5a) cannot recapitulate normal *Xist* expression[7].

The respective promoters of *Xist* and *Tsix* lie in two neighbouring TADs with transcription crossing the intervening boundary (Fig. 2b), consistent with previous 3C experiments[22]. Whereas the *Xist* promoter and its positive regulators are located in TAD E, the promoter of its antisense repressor, *Tsix*, lies in TAD D, which extends up to *Ppnx* (also known as *4930519F16Rik*)/*Nap1l2*, more than 200 kb away (Fig. 2b). Thus, in addition to the *Xite* enhancer, more distant elements within TAD D may participate in *Tsix* regulation. To test this we used two different single-copy transgenic mouse lines, Tg53 and Tg80 (ref. 23). Both transgenes contain *Xist*, *Tsix* and *Xite* (Fig. 5a). Tg53 encompasses the whole of TAD D, whereas Tg80 is truncated just 5′ to *Xite* (Fig. 5a and Supplementary Fig. 9). In the inner cell mass of male mouse embryos at embryonic day 4.0 (E4.0), *Tsix* transcripts could be readily detected from Tg53, as well as from the endogenous X (Fig. 5b). However, no *Tsix* expression could be detected from Tg80, which lacks the distal portion of TAD D (Fig. 5b). Thus, sequences within TAD D must contain essential elements for the correct developmental regulation of *Tsix*.

Within TAD D, several significant looping events involving the *Tsix* promoter or its enhancer *Xite* were detected (Figs 2b and 5a, Supplementary Fig. 10). Alignment of 5C maps with chromatin signatures of enhancers in mouse ESCs (Supplementary Fig. 11) suggested the existence of multiple regulatory elements within this region. We also identified a transcript initiating approximately 50 kb upstream of the *Ppnx* promoter (Fig. 5a), from a region bound by pluripotency factors and corresponding to a predicted promoter for a large (80 kb) intervening non-coding RNA (lincRNA[24], Supplementary Fig. 12) which we termed *Linx* (large intervening transcript in the *Xic*). *Linx* RNA shares several features with non-coding RNAs, such as accumulation around its transcription site[25] (Fig. 5c), nuclear enrichment and abundance of the unspliced form[26] (Supplementary Fig. 12 and 13). *Linx* and *Tsix* are co-expressed in the inner cell mass of blastocysts from E3.5–4.0 onwards, as well as in male and female mouse ESCs (Fig. 5c). *Linx* RNA is not detected earlier in embryogenesis, nor in extra-embryonic lineages, implying an epiblast-specific function

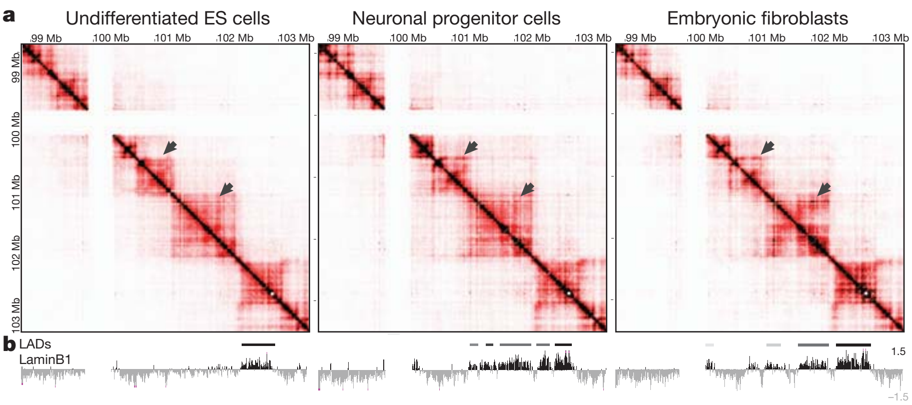


Figure 3 | Dynamics of topologically associating domains during cell differentiation. **a**, Comparison of 5C data from male mouse ESCs (E14), NPCs (E14) and primary MEFs reveals general conservation of TAD positions during differentiation, but differences in their internal organization (arrows highlight examples of tissue-specific patterns). **b**, Lamina-associated domains (LADs, from ref. 19) align with TADs. Chromosomal positions of tissue-specific LADs reflect gain of lamina association by TADs, as well as internal reorganization of lamina-associated TADs during differentiation.

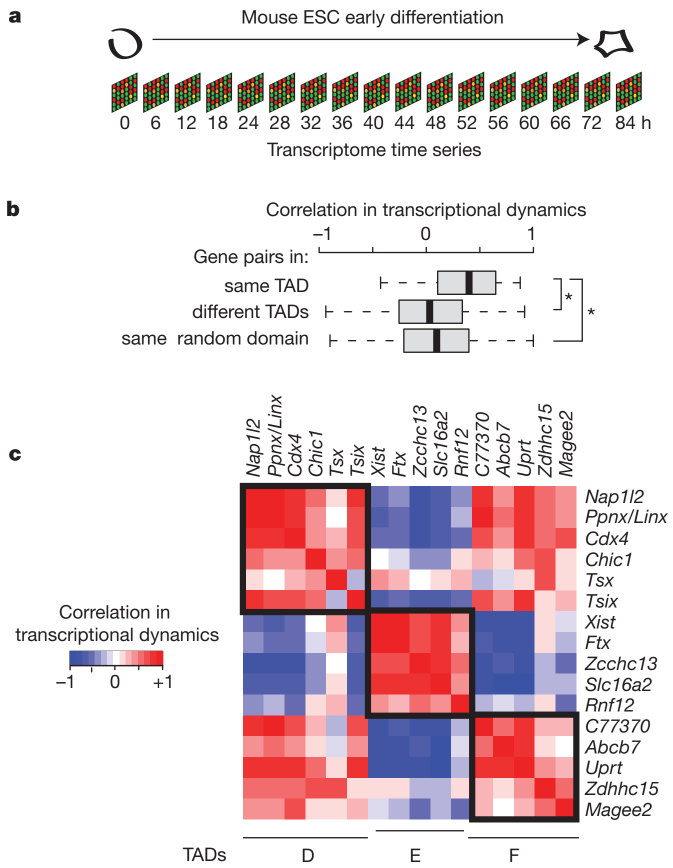


Figure 4 | Transcriptional co-regulation within topologically associating domains. **a**, Female mouse ESCs were differentiated towards the epiblast stem cell lineage for 84 h. Transcript levels were measured every 4–6 h at 17 different time points by microarray analysis. **b**, Pearson's correlation coefficients over all time points were calculated for gene pairs lying in the same TAD, pairs in different TADs and for pairs in randomly defined domains on the X chromosome that contain a similar number of genes and are of comparable size. Boxplots show the distribution of Pearson's correlation coefficients, with whiskers and boxes encompassing all and 50% of values, respectively, and central bars denoting the median correlation coefficient. * represents significant difference with $P < 10^{-7}$ using Wilcoxon's rank sum test. **c**, Pearson's correlation coefficients for gene pairs in TADs D, E and F with red denoting positive and blue negative correlation. Boxes indicate the TAD boundaries.

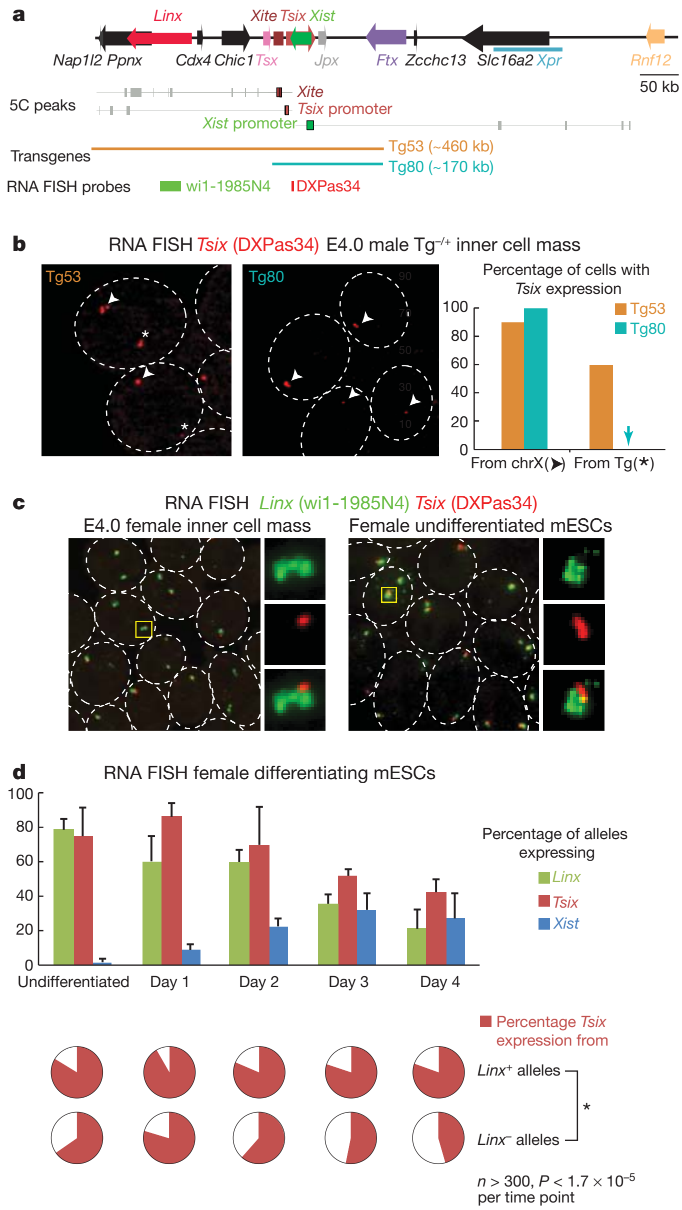


Figure 5 | 5C maps reveal new regulatory regions in the *Xic*. **a**, Statistically significant looping events (5C peaks) for restriction fragments within *Xite*, *Tsix* promoter or *Xist* promoter within their respective TAD, in male (E14) mouse ESCs. The Tg80 YAC transgene lacks genomic elements found to interact physically with *Xite*/*Tsix* that are present in Tg53. **b**, RNA FISH analysis of *Tsix* expression is detected in the inner cell masses of heterozygous transgenic male E4.0 embryos by RNA FISH from single-copy paternally inherited Tg53 but not Tg80 transgenes. Transgenic (star) and endogenous *Tsix* alleles (arrowhead) were discriminated by subsequent DNA FISH as in Supplementary Fig. 5. $n = 20$ inner cell mass cells (two embryos each). **c**, *Linx* transcripts (green, wi1-1985N4 probe) are expressed in both E4.0 inner cell mass cells and mouse ESCs, together with *Tsix* (red, DXPas34 probe), and unspliced transcripts accumulate locally in a characteristic cloud-like shape. **d**, RNA FISH in differentiating female mouse ESCs revealing synchronous downregulation of *Linx* and *Tsix* with concomitant upregulation of *Xist* (detected with a strand-specific probe). Bars are the standard deviation around the mean of three experiments. Triple-colour RNA FISH allows simultaneous detection of *Linx*, *Tsix* and *Xist* RNAs. Scoring of *Xist*-negative alleles demonstrates that before *Xist* upregulation *Tsix* expression is more frequent from *Linx*-expressing alleles than from *Linx* non-expressing alleles, at all time points tested. Presented is the mean of three experiments. Statistical differences were assessed using Fisher's exact test. Cells were differentiated in monolayers by withdrawal of leukaemia inhibitory factor (LIF).

(Supplementary Fig. 9). Triple RNA FISH for *Linx*, *Tsix* and *Xist* in differentiating female mouse ESCs (Supplementary Fig. 14) revealed that before *Xist* upregulation, the probability of *Tsix* expression from alleles co-expressing *Linx* is significantly higher than from alleles that do not express *Linx* (Fig. 5d). Furthermore, *Linx* expression is frequently monoallelic, even before *Xist* upregulation (Supplementary Fig. 14), revealing a transcriptional asymmetry of the two *Xic* alleles before XCI. Taken together, our experiments based on 5C, transgenesis and RNA FISH, point towards a role for *Linx* in the long-range transcriptional regulation of *Tsix* — either through its chromosomal association with *Xite* and/or via the RNA it produces. This analysis of the *Xist*/*Tsix* region illustrates how spatial compartmentalization of chromosomal neighbourhoods in TADs partitions the *Xic* into two large regulatory domains, with opposite transcriptional fates (Supplementary Fig. 15).

In conclusion, our study reveals that sub-megabase folding of mammalian chromosomes results in the self-association of large chromosomal neighbourhoods in the three-dimensional space of the nucleus. The stability of such partitioning throughout differentiation, X inactivation and in cell lines with impaired histone-modifying machineries, indicates that this level of chromosomal organization may provide a basic framework onto which other domain-wide

features, such as lamina association and blocks of histone modification, can be dynamically overlaid. Our data also point to a role for TADs in shaping regulatory landscapes, by defining the extent of sequences that belong to the same regulatory neighbourhood. We anticipate that TADs may underlie regulatory domains previously proposed on the basis of functional and synteny conservation studies[27,28]. We believe that the principles we have revealed here will not be restricted to the *Xic*, as spatial partitioning of chromosomal neighbourhoods occurs throughout the genome of mouse and human[29], as well as *Drosophila*[30] and *E. coli*[31]. We have shown that TAD boundaries can have a critical role in high-order chromatin folding and proper long-range transcriptional control. Future work will clarify the mechanisms driving this level of chromosomal organization, and to what extent it generally contributes to transcriptional regulation. In summary, our study provides new insights into the *cis*-regulatory architecture of chromosomes that orchestrates transcriptional dynamics during development, and paves the way to dissecting the constellation of control elements of *Xist* and its regulators within the *Xic*.

METHODS SUMMARY

5C was performed on mouse ESCs, mouse NPCs and primary MEFs following a previously described protocol[2] with modifications, and sequenced on one lane of an Illumina GAIIx. RNA and DNA FISH were performed on mouse ESCs and inner cell masses extracted from pre-implantation embryos as previously described[7], with modifications. Full experimental and bioinformatic methods are detailed in Supplementary Information.

Received 3 October 2011; accepted 22 March 2012.
Published online 11 April 2012.

1. Kleinjan, D. A. & Lettice, L. A. Long-range gene control and genetic disease. *Adv. Genet.* **61,** 339–388 (2008).
2. Dostie, J. *et al.* Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16,** 1299–1309 (2006).
3. Rastan, S. Non-random X-chromosome inactivation in mouse X-autosome translocation embryos–location of the inactivation centre. *J. Embryol. Exp. Morphol.* **78,** 1–22 (1983).
4. Rastan, S. & Robertson, E. J. X-chromosome deletions in embryo-derived (EK) cell lines associated with lack of X-chromosome inactivation. *J. Embryol. Exp. Morphol.* **90,** 379–388 (1985).
5. Augui, S., Nora, E. P. & Heard, E. Regulation of X-chromosome inactivation by the X-inactivation centre. *Nature Rev. Genet.* **12,** 429–442 (2011).
6. Anguera, M. C. *et al. Tsx* produces a long noncoding RNA and has general functions in the germline, stem cells, and brain. *PLoS Genet.* **7,** e1002248 (2011).
7. Heard, E., Mongelard, F., Arnaud, D. & Avner, P. *Xist* yeast artificial chromosome transgenes function as X-inactivation centers only in multicopy arrays and not as single copies. *Mol. Cell. Biol.* **19,** 3156–3156 (1999).
8. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
9. Marks, H. *et al.* High-resolution analysis of epigenetic changes associated with X inactivation. *Genome Res.* **19,** 1361–1373 (2009).
10. Pauler, F. M. *et al.* H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res.* **19,** 221–233 (2009).
11. Wen, B., Wu, H., Shinkai, Y., Irizarry, R. A. & Feinberg, A. P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nature Genet.* **41,** 246–250 (2009).
12. Lienert, F. *et al.* Genomic prevalence of heterochromatic H3K9me2 and transcription do not discriminate pluripotent from terminally differentiated cells. *PLoS Genet.* **7,** e1002090 (2011).
13. Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6,** 479–491 (2010).
14. Rougeulle, C. *et al.* Differential histone H3 Lys-9 and Lys-27 methylation profiles on the X chromosome. *Mol. Cell. Biol.* **24,** 5475–5484 (2004).
15. Montgomery, N. D. *et al.* The murine polycomb group protein Eed is required for global histone H3 lysine-27 methylation. *Curr. Biol.* **15,** 942–947 (2005).
16. Monkhorst, K., Jonkers, I., Rentmeester, E., Grosveld, F. & Gribnau, J. X Inactivation counting and choice is a stochastic process: evidence for involvement of an X-linked activator. *Cell* **132,** 410–421 (2008).
17. Spencer, R. J. *et al.* A boundary element between *Tsix* and *Xist* binds the chromatin insulator Ctcf and contributes to initiation of X chromosome inactivation. *Genetics* CrossRef (2011).
18. Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467,** 430–435 (2010).
19. Peric-Hupkes, D. *et al.* Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol. Cell* **38,** 603–613 (2010).
20. Splinter, E. *et al.* The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev.* **25,** 1371–1383 (2011).
21. Caron, H. *et al.* The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291,** 1289–1292 (2001).
22. Tsai, C.-L., Rowntree, R. K., Cohen, D. E. & Lee, J. T. Higher order chromatin structure at the X-inactivation center via looping DNA. *Dev. Biol.* **319,** 416–425 (2008).
23. Heard, E. *et al.* Transgenic mice carrying an *Xist*-containing YAC. *Hum. Mol. Genet.* **5,** 441–450 (1996).
24. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–227 (2009).
25. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA* **106,** 11667–11672 (2009).
26. Seidl, C. I. M., Stricker, S. H. & Barlow, D. P. The imprinted *Air* ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. *EMBO J.* **25,** 3565–3575 (2006).
27. Ruf, S. *et al.* Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nature Genet.* **43,** 379–386 (2011).
28. Kikuta, H. *et al.* Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17,** 545–555 (2007).
29. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* doi:10.1038/nature11082 (this issue).
30. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148,** 458–472 (2012).
31. Mercier, R. *et al.* The MatP/*matS* site-specific system organizes the terminus region of the *E. coli* chromosome into a macrodomain. *Cell* **135,** 475–485 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Pollex and T. Forné for experimental help; the imaging facility PICTIBiSA@BDD for technical assistance, D. Gentien and C. Hego for microarray hybridizations. We thank K. Bernhard, F. Stewart and A. Smith for protocols and material for 2i culture and EpiSC differentiation. We are grateful to members of the E.H. laboratory for critical input. This work was funded by grants from the Ministère de la Recherche et de l'Enseignement Supérieur and the ARC (to E.P.N.); a HFSP Long term fellowship (LT000597/2010-L) (to E.G.S.). EU EpiGeneSys FP7 Network of Excellence no. 257082, the Fondation pour la Recherche Medicale, ANR, ERC Advanced Investigator award no. 250367 and EU FP7 SYBOSS grant no. 242129 (to E.H.). N.B. was supported by BMBF (FORSYS) and EMBO (fellowship ASTF 307-2011). J.D., B.R.L. and N.L.v.B. were supported by NIH (R01 HG003143) and a W. M. Keck Foundation Distinguished Young Scholar Award.

Author Contributions E.P.N. performed and analysed 3C, 5C, (RT–)qPCR, immunofluorescence, RNA and DNA FISH. B.R.L. and N.L.v.B. helped in the design and/or the analysis of 3C and 5C. L.G. performed 3C, FISH and 5C analysis. E.G.S. generated the time-course transcriptomic data, which was analysed by J.M. and N.B.; I.O. performed FISH on pre-implantation embryos. J.G. donated the XTX mouse ESC line. N.S. and E.B. helped in the epigenomic and 5C analyses. J.S. and T.P. set up OMX microscopy and analysis and T.P. performed structured illumination microscopy and image analysis. The manuscript was written by E.P.N., J.D. and E.H. with contribution from E.G.S. and input from all authors.

Author Information High-throughput data are deposited in Gene Expression Omnibus under accession number GSE35721 for all 5C experiments and GSE34243 for expression microarrays. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.H. (edith.heard@curie.fr) or J.D. (job.dekker@umassmed.edu).

LETTER

doi:10.1038/nature11070

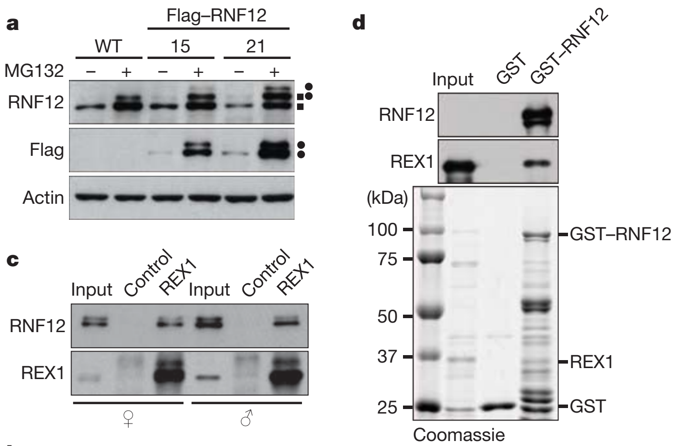
RNF12 initiates X-chromosome inactivation by targeting REX1 for degradation

Cristina Gontan[1], Eskeatnaf Mulugeta Achame[1], Jeroen Demmers[2], Tahsin Stefan Barakat[1], Eveline Rentmeester[1], Wilfred van IJcken[3], J. Anton Grootegoed[1] & Joost Gribnau[1]

Evolution of the mammalian sex chromosomes has resulted in a heterologous X and Y pair, where the Y chromosome has lost most of its genes. Hence, there is a need for X-linked gene dosage compensation between XY males and XX females. In placental mammals, this is achieved by random inactivation of one X chromosome in all female somatic cells[1]. Upregulation of *Xist* transcription on the future inactive X chromosome acts against *Tsix* antisense transcription, and spreading of *Xist* RNA in *cis* triggers epigenetic changes leading to X-chromosome inactivation. Previously, we have shown that the X-encoded E3 ubiquitin ligase RNF12 is upregulated in differentiating mouse embryonic stem cells and activates *Xist* transcription and X-chromosome inactivation[2]. Here we identify the pluripotency factor REX1 as a key target of RNF12 in the mechanism of X-chromosome inactivation. RNF12 causes ubiquitination and proteasomal degradation of REX1, and *Rnf12* knockout embryonic stem cells show an increased level of REX1. Using chromatin immunoprecipitation sequencing, REX1 binding sites were detected in *Xist* and *Tsix* regulatory regions. Overexpression of REX1 in female embryonic stem cells was found to inhibit *Xist* transcription and X-chromosome inactivation, whereas male *Rex1*$^{+/-}$ embryonic stem cells showed ectopic X-chromosome inactivation. From this, we propose that RNF12 causes REX1 breakdown through dose-dependent catalysis, thereby representing an important pathway to initiate X-chromosome inactivation. *Rex1* and *Xist* are present only in placental mammals, which points to co-evolution of these two genes and X-chromosome inactivation.

The initiation of X-chromosome inactivation (XCI) exclusively in female cells implies a need for X-linked XCI activators that act in a dose-dependent manner to sense the number of X chromosomes present per diploid genome[3,4]. We recently identified X-encoded RNF12 as a dose-dependent activator of XCI in mouse embryonic stem cells (ESCs)[2]. Additional transgenic copies of *Rnf12* resulted in initiation of XCI in male cells, and on both X chromosomes in a high percentage of female cells[2]. Random XCI was found to be markedly reduced in differentiating *Rnf12*$^{+/-}$ and *Rnf12*$^{-/-}$ female ESCs, which indicated an important role for RNF12 in the regulation of XCI, although the mechanism by which this E3 ubiquitin ligase initiates XCI remained elusive[5,6].

To address this question, we generated Flag–*Rnf12* transgenic female *Rnf12*$^{+/-}$ ESCs to identify interaction partners of RNF12 by Flag-affinity purification. RNF12 is very unstable and the addition of the proteasome inhibitor MG132 facilitates its detection (Fig. 1a). Flag–RNF12 was purified from nuclear extracts of two Flag–*Rnf12* ESC lines (Supplementary Fig. 1a). Purified RNF12 samples and control samples were separated on SDS polyacrylamide gels (Supplementary Fig. 1b) and analysed by mass spectrometry (Supplementary Table 1). The only transcription factor consistently co-purifying with RNF12 was REX1 (Fig. 1b and Supplementary Table 1). Previous studies have demonstrated that *Rex1* expression strictly correlates with the pluripotent state of ESCs[7], and REX1 has been implicated in suppression of genes involved in ESC differentiation[8]. We performed the reverse experiment, using two transgenic female ESC lines expressing a Flag–V5-tagged REX1 fusion protein (Supplementary Fig. 1c). In both REX1 purifications (Supplementary Fig. 1d, e), performed with nuclear extracts of undifferentiated transgenic ESCs, RNF12 was present as a prominent interacting partner (Fig. 1b). Co-purified RNF12 was non-ubiquitinated, but mass spectometry analysis and phosphatase treatment indicated that a significant fraction of RNF12 is phosphorylated (Supplementary Fig. 1f). We confirmed the REX1–RNF12 interaction by co-immunoprecipitation of endogenous RNF12 and REX1 from undifferentiated female and male ESCs nuclear extracts (Fig. 1c), and co-immunoprecipitation of recombinant glutathione *S*-transferase



b

	Flag purifications							
	Flag–RNF12_15		Flag–RNF12_21		Flag–V5–REX1_3		Flag–V5–REX1_11	
Protein	Mascot	Peptides	Mascot	Peptides	Mascot	Peptides	Mascot	Peptides
RNF12	1308	21(40)	1490	23(48)	1214	18(95)	1535	22(132)
REX1	124	2(5)	61	1(4)	558	11(71)	603	11(89)

Figure 1 | RNF12 interacts with REX1 in mouse ESCs. a, Nuclear extracts of wild-type (WT) and transgenic Flag–*Rnf12* ESC clones 15 and 21 were immunoblotted with RNF12 and Flag antibodies (running positions of WT RNF12 and Flag–RNF12 are indicated by filled squares and circles). Where indicated, cells were treated with proteasome inhibitor (MG132). Actin was used as a loading control. **b**, Mass spectrometry analysis of Flag-affinity purifications from Flag–RNF12-expressing ESC clones 15 and 21, and Flag–REX1-expressing clones 3 and 11. Mascot score, number of unique peptides and total number of peptides identified (between brackets) are shown for RNF12 and REX1 proteins in each of the purifications. **c**, REX1–RNF12 co-immunoprecipitation from nuclear extracts of female (left) and male (right) ESCs. Immunoprecipitations with REX1 antibody or control rabbit IgG were immunoblotted with RNF12 and REX1 antibodies. **d**, Direct binding of recombinant GST–RNF12 to recombinant REX1. RNF12 and REX1 are detected by immunoblotting (upper panels) and by Coomassie staining of an SDS–polyacrylamide gel electrophoresis gel (lower panel). GST alone was used as a negative control.

[1]Department of Reproduction and Development, Erasmus MC, University Medical Center, Dr Molewaterplein 50, 3015 GE Rotterdam, The Netherlands. [2]Proteomics Center, Erasmus MC, University Medical Center, Dr Molewaterplein 50, 3015 GE Rotterdam, The Netherlands. [3]Biomics Department, Erasmus MC, University Medical Center, Dr Molewaterplein 50, 3015 GE Rotterdam, The Netherlands.

(GST)–RNF12 and REX1 (Fig. 1d). Mapping of the RNF12 region(s) involved in the interaction with REX1 indicated that both the amino (N)- and carboxy (C)-terminal halves of RNF12 contribute to the interaction (Supplementary Fig. 2a, b).

We generated two catalytically inactive RNF12 mutants (Fig. 2a). Transient expression in ESCs showed an increased stability of the RNF12 mutants (Supplementary Fig. 2c), indicating that autoubiquitination contributes to the high turnover of RNF12. To test whether REX1 is a bona fide substrate of RNF12, we transfected combinations of expression vectors encoding wild-type or mutant RNF12–green fluorescent protein (GFP), with REX1–Cherry fusion proteins into HEK293 cells (Fig. 2b and Supplementary Fig. 3a, b). Fluorescence-activated cell sorting analysis of REX1–Cherry transfected HEK 293 cells showed a more than tenfold decrease in Cherry intensity when cells were co-transfected with wild-type RNF12 compared with co-transfection with the RNF12 mutants (Fig. 2c and Supplementary Fig. 3c, d). This result suggests a strict correlation between RNF12 expression and REX1 degradation.

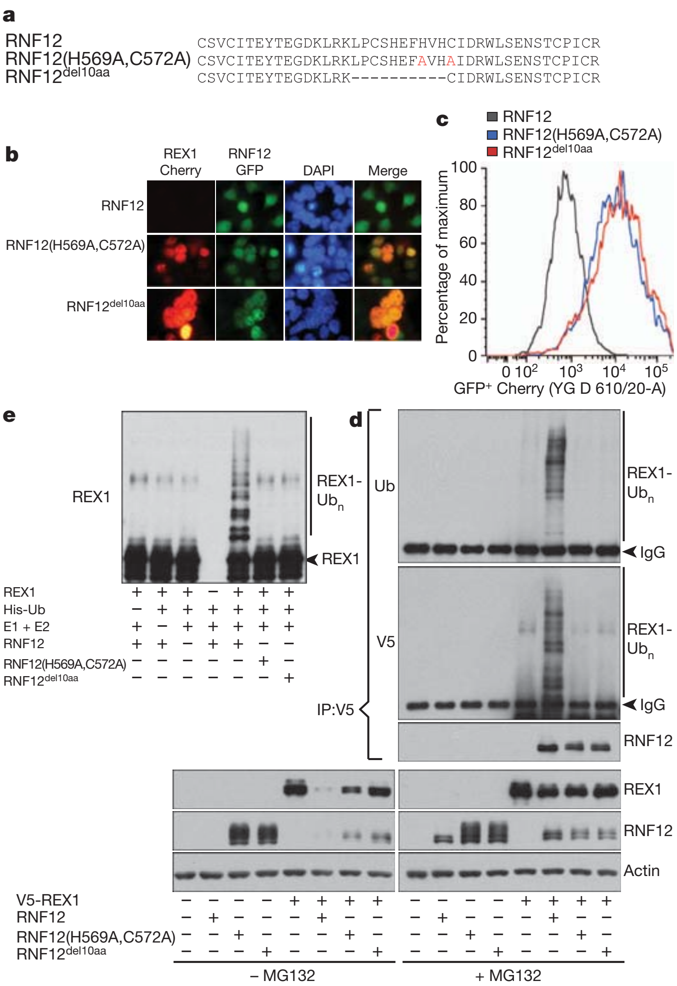


Figure 2 | RNF12 polyubiquitinates and targets REX1 for proteasomal degradation. **a**, Sequence alignment of the RING-finger domain corresponding to the WT and two catalytically inactive RNF12 mutants (two arginine substitutions are shown in red). **b**, Analysis of the fluorescence in HEK293 cells, 24 h after transfection with REX1–Cherry fusion expression vector alone or in combination with WT or catalytically inactive RNF12–GFP expression constructs. **c**, Quantification of **b** by fluorescence-activated cell sorting analysis. **d**, Left bottom panels: nuclear protein extracts of HEK293 cells co-transfected with the indicated expression constructs for V5-REX1, WT RNF12 or the catalytically inactive RNF12 mutants. Immunoblots were probed with the indicated antibodies. Right bottom panels: the same transfections, but in the presence of MG132. Upper panels: V5-tagged REX1 was immunoprecipitated with anti-V5 agarose beads and analysed by immunoblotting to detect the poly-ubiquitinated REX1 (REX1-Ub$_n$) with V5 and ubiquitin (Ub) antibodies. RNF12 co-immunoprecipitated with REX1 was detected with RNF12 antibody. **e**, REX1 ubiquitination assay with recombinant proteins. Recombinant REX1, His-ubiquitin, UBE1 (E1), UBCH5A (E2), recombinant RNF12 and RNF12 mutants were added in the indicated combinations. Ubiquitinated forms of REX1 were detected by immunoblotting with REX1 antibody.

To provide evidence that REX1 is indeed ubiquitinated by RNF12, HEK293 cells were transfected with V5-REX1, and either RNF12 wild-type or the two inactive RNF12 mutants. REX1 was degraded in the presence of wild-type RNF12 but not in the presence of the RNF12 mutants (Fig. 2d), and degradation was blocked by the proteasome inhibitors MG132 or epoxomicin (Fig. 2d and Supplementary Fig. 4). We subjected the nuclear extracts to immunoprecipitation with anti-V5 agarose beads, and probed with V5 and ubiquitin antibodies to visualize poly-ubiquitinated REX1 (Fig. 2d and Supplementary Fig. 4). Mass spectrometric analysis detected five putative lysine acceptor sites for ubiquitin linkage (Supplementary Fig. 5). In addition, an ubiquitination assay performed with recombinant proteins revealed poly-ubiquitination of REX1 only in the presence of wild-type RNF12, but not in the presence of either of the two RNF12 mutants (Fig. 2e). These results indicate a direct role for RNF12 in targeting REX1 for degradation by the proteasome in an ubiquitin-dependent manner.

Our results predict that RNF12 and REX1 protein levels show a reciprocal correlation. Indeed, we found a pronounced increase of the REX1 protein level in *Rnf12*$^{-/-}$ ESCs (Fig. 3a). Western blot analysis of RNF12 and REX1 protein levels in female wild-type ESCs indicated that REX1 is quickly downregulated upon differentiation, coinciding with an initial increase in RNF12 (Fig. 3b). In *Rnf12*$^{-/-}$ ESCs, the REX1 protein level is also downregulated but the starting level is much higher. Comparison of the REX1 protein level in wild-type, *Rnf12*$^{+/-}$ and *Rnf12*$^{-/-}$ female ESCs, and in wild-type male cells, by immunoblotting, indicates that the REX1 level is very low in wild-type female ESCs (Supplementary Fig. 6a), in agreement with a high ubiquitination-dependent turnover of REX1. Female *Rnf12*$^{+/-}$ cells and wild-type male cells display a lower RNF12 expression level and an increased REX1 protein level, although this relationship was not strictly twofold, which may be related to the pluripotent state of the ESC lines and enzyme kinetics (Supplementary Fig. 6a and Fig. 3c). *Rnf12* gene dosage did not affect other known factors involved in XCI, including YY1, NANOG and SUZ12 (Fig. 3b and Supplementary Fig. 6a). *Rex1* gene transcription, analysed by quantitative PCR (qPCR), was not affected in *Rnf12*$^{-/-}$ cells (Supplementary Fig. 6b), indicating that the downregulation of REX1 by RNF12 is post-transcriptional.

The marked increase of REX1 protein level in *Rnf12*$^{-/-}$ cells points to an increase in REX1 protein stability in the absence of RNF12. Indeed, half-life experiments, after cycloheximide treatment of V5-Rex1 transfected wild-type and *Rnf12*$^{-/-}$ ESCs, indicated a strong increase in the half life ($t_{1/2}$) of REX1 in *Rnf12*$^{-/-}$ ESCs ($t_{1/2} > 2$ h) compared with wild-type cells ($t_{1/2} < 0.5$ h) (Fig. 3d). Also, the REX1 protein level was downregulated more effectively with an increasing dose of RNF12, as detected in HEK293 cells co-transfected with a fixed concentration of V5-*Rex1* and an increasing concentration of Flag–*Rnf12* expression vectors (Fig. 3e). We previously found that male (as well as female) ESC lines stably overexpressing *Rnf12* show ectopic XCI[2]. To determine whether overexpression of *Rnf12* would also affect the REX1 level, we stably introduced a bacterial artificial chromosome covering *Rnf12* into male ESCs, resulting in ectopic XCI at day 3 of differentiation (Supplementary Fig. 6c). When undifferentiated, these clones showed a lower level of REX1, which confirms that overexpression of RNF12 in male cells lowers REX1 stability (Supplementary Fig. 6c). Analysis by qPCR showed that *Xist* is upregulated in ESCs transiently overexpressing wild-type RNF12, which is not observed using mutant RNF12, demonstrating that an intact RING-finger is required for initiation of XCI (Supplementary Fig. 6d).

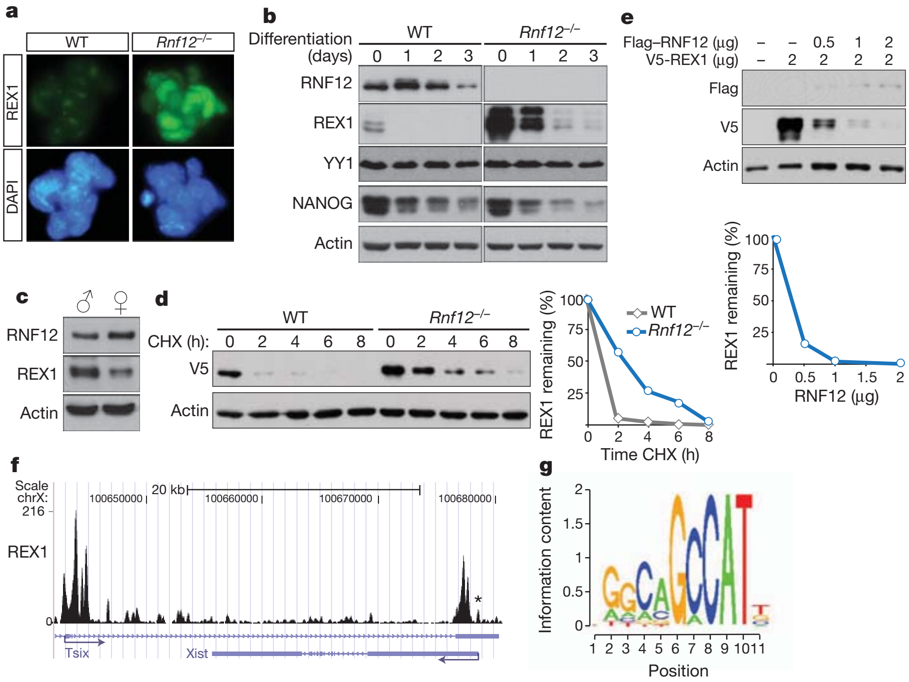


Figure 3 | RNF12 is a dose-dependent regulator of REX1 expression. **a**, REX1 immunostaining (green) on WT and *Rnf12*$^{-/-}$ ESCs. **b**, Immunoblots of WT and *Rnf12*$^{-/-}$ ESCs at day 0, 1, 2 and 3 of differentiation, probed with antibodies against RNF12, REX1, YY1 and NANOG. **c**, RNF12 and REX1 protein levels were detected in female and male ESC nuclear extracts by immunoblotting. **d**, REX1 half-life measurements in WT and *Rnf12*$^{-/-}$ ESCs. ESCs were transiently transfected with V5-REX1 and treated 24 h after transfection with 100 μg ml^{-1} cycloheximide (CHX) for the indicated times. Left panel: nuclear protein extracts were immunoblotted for V5-REX1 with V5 antibody. Right panel: quantification of the REX1 level, using ImageJ software, at different time points compared with $t = 0$ h (100%) in WT or *Rnf12*$^{-/-}$ ESCs. **e**, REX1 degradation by RNF12 is dose dependent. Flag–*Rnf12* or V5-*Rex1* constructs were co-transfected into HEK293 cells. Upper panel: levels of V5-REX1 and Flag–RNF12 visualized by immunoblotting with V5 or Flag antibodies. Bottom panel: quantification of the V5-REX1 level in cells co-transfected with RNF12, compared with the 100% level in cells transfected only with V5-REX1 plasmid. Actin was used as a loading control in **b**–**e**. **f**, REX1 binding pattern in the *Xist*/*Tsix* genomic region in female ESCs, as determined by V5-REX1 ChIP-seq. Identified sequence reads were plotted relative to genomic location and visualized using the University of California, Santa Cruz (UCSC) Genome Browser. Location and transcription start sites (arrows) of the *Tsix* and *Xist* loci are indicated. The asterisk marks a REX1 binding site in the *Xist* promoter. **g**, The REX1 consensus motif highly enriched in the genome-wide REX1 ChIP-seq peaks.

For RNF12 to function as an XCI-activator through degradation of REX1, REX1 could either repress *Xist* or activate *Tsix*. Indeed, a recent chromatin immunoprecipitation (ChIP)–qPCR study identified REX1 recruitment to the *Tsix* regulatory element *DXPas34*, which was found to be important for effective elongation of RNA polymerase II[9]. To identify all binding sites of REX1 in the region encompassing *Xist* and *Tsix*, we performed ChIP-sequencing (ChIP-seq) analysis on undifferentiated Flag–V5-*Rex1* female ESCs. This analysis confirmed enrichment for previously published REX1 binding sequences in *Tsix*[9], and showed specific REX1 binding sites in the *Xist* promoter and promoter distal region (Fig. 3f), although REX1 recruitment to the *Xist* promoter was detected only in the presence of MG132. The observed genome-wide REX1 peaks revealed a highly enriched consensus-binding motif (Fig. 3g). Recruitment of REX1 to *Xist* and *Tsix* was also detectable, but less prominent, in the absence of MG132 (Supplementary Fig. 7). This result may explain why REX1 recruitment to *Xist* was not detected in the previous study[9], which did not include the use of proteasome inhibitors. Our results indicate that REX1 may perform a dual function, in the repression of *Xist* and the activation of *Tsix*.

To elucidate the role of REX1 in XCI, we analysed XCI in day-3-differentiated *Rex1*$^{+/-}$ male ESCs[10] and control male ESCs. As expected, we found a small percentage of male control cells that contained *Xist* clouds (1%), detected with *Xist* RNA fluorescence *in situ* hybridization (FISH). In contrast, as many as 7.5% of the *Rex1*$^{+/-}$ male cells showed *Xist* clouds, hence initiation of XCI, supporting a dose-dependent role for REX1 in repression of XCI (Fig. 4a, b). We next analysed the female ESC lines overexpressing Flag–V5-tagged REX1 and determined the percentage of cells that initiated XCI after 3 days of differentiation, by *Xist* RNA FISH. For both REX1 overexpressing lines (Rex1_3 and Rex1_11) we detected a severely reduced number of cells with an *Xist* coated inactive X chromosome (Xi), compared with wild-type female cells (Fig. 4c, d), indicating a strong inhibition of XCI. *Rnf12*$^{-/-}$ cells showed no *Xist* clouds, consistent with our previous studies[6]. To test whether the XCI phenotype in *Rnf12*$^{-/-}$ ESCs was directly related to the resulting increased REX1 level we performed *Rex1* knockdown experiments in *Rnf12*$^{-/-}$ ESCs. A reduction in *Rex1* expression by more than 50%, 3 days after transient transfection of a *Rex1* short hairpin RNA (shRNA) vector, resulted in a more than fivefold induction of *Xist*, in agreement with a mechanism in which REX1 acts downstream of RNF12 to activate XCI (Fig. 4e).

Next we analysed the RNA expression level of *Rex1*, *Xist* and *Tsix* by qPCR in undifferentiated and day-3-differentiated wild-type and *Rex1* overexpressing ESC lines. In the last group of cells, *Rex1* mRNA expression was 2.8- to 4.5-fold upregulated before differentiation, followed by partial downregulation at day 3 of differentiation owing to silencing of endogenous *Rex1* and reduced expression of the transgene (Fig. 4f). Analysis by qPCR of differentiation and pluripotency markers provided evidence that the *Rex1* overexpressing ESC lines undergo proper differentiation, which implies that the XCI phenotype is not a consequence of a differentiation defect (Supplementary Fig. 8). Our analysis also indicated that *Tsix* expression was slightly upregulated in undifferentiated REX1 overexpressing compared with wild-type cells. Expression of *Tsix* was downregulated at day 3 of differentiation, but was still higher than in wild-type cells, consistent

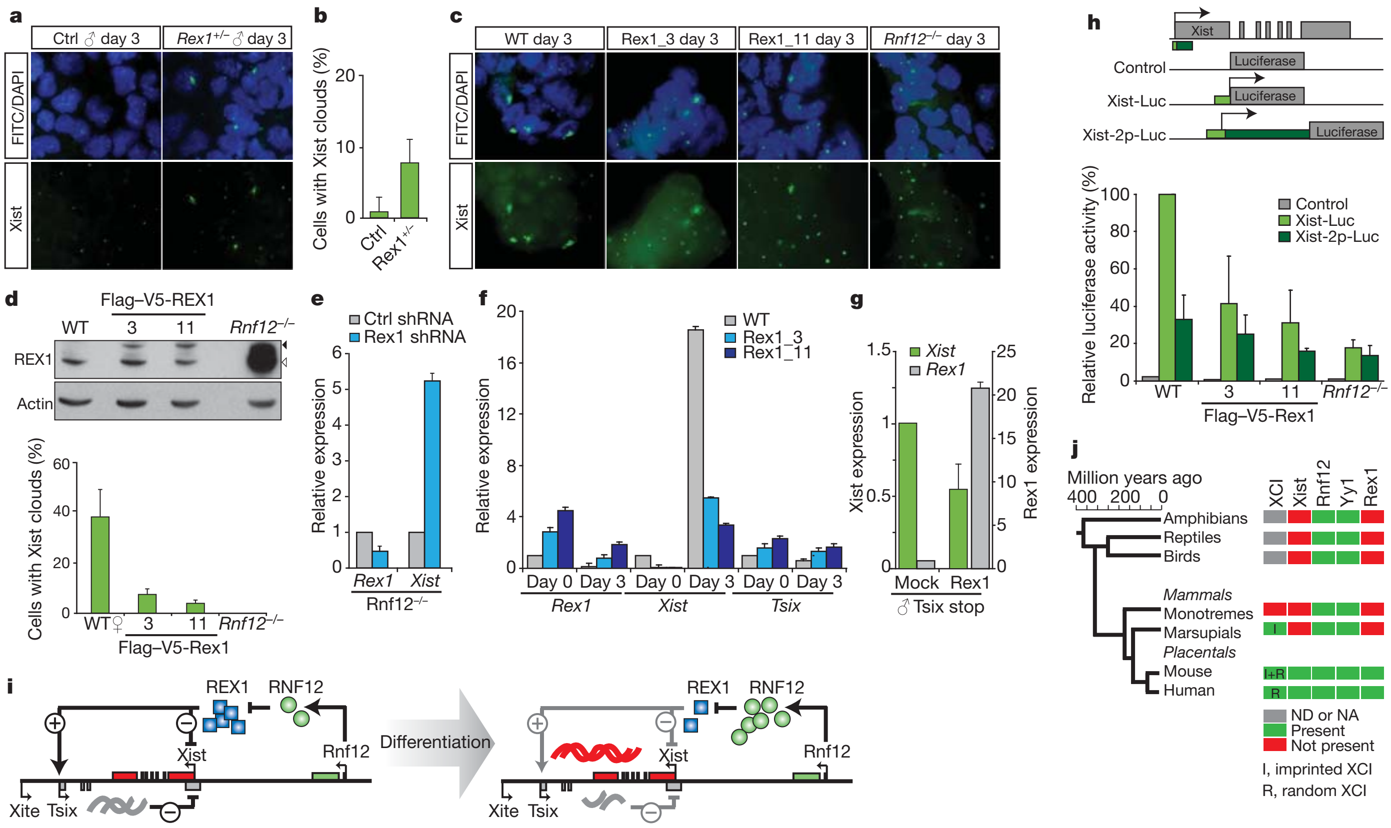


Figure 4 | REX1-dependent regulation of XCI. **a**, *Xist* RNA-FISH (fluorescein isothiocyanate (FITC)) on day-3-differentiated WT and *Rex1*$^{+/-}$ male ESCs. **b**, Quantification of *Xist* clouds in **a** ($n > 100$; error bars, 95% Wilson confidence interval). **c**, *Xist* RNA-FISH (FITC) on day-3-differentiated female WT, *Rex1* overexpressing clones 1_3 and 1_11, and *Rnf12*$^{-/-}$ ESCs. **d**, Top panel: nuclear extracts of WT, transgenic Flag–V5-REX1 lines 3 and 11 and *Rnf12*$^{-/-}$ ESCs were immunoblotted with REX1 antibody (open triangle, endogenous REX1; filled triangle, Flag–V5-REX1). Actin was used as a loading control. Bottom panel: quantification of *Xist* clouds in **c** ($n > 200$). Average percentage of cells with *Xist* clouds is shown. **e**, *Xist* and *Rex1* qPCR analysis of *Rnf12*$^{-/-}$ ESCs 72 h after transfection with a *Rex1* shRNA construct or a control vector. **f**, qPCR analysis of *Rex1*, *Xist* and *Tsix* expression in undifferentiated and day-3-differentiated WT and the Rex1_3 and Rex1_11 clones. **g**, qPCR analysis of *Rex1* and *Xist* expression in day-3-differentiated male *Tsix*-stop ESCs, after transient transfection with a *Rex1* or control (mock) expression vector. **h**, Upper panel: representation of constructs used in the luciferase reporter assay—empty vector, *Xist* promoter (Xist-luc) and *Xist* promoter + proximal part of exon 1 (Xist-2p-luc) constructs. Lower panel: luciferase activity of the different constructs in WT, Rex1_3 and Rex1_11 clones and *Rnf12*$^{-/-}$ female ESCs, tranfected with the corresponding reporter constructs and differentiated for 3 days. All data in **d**–**h** represent the average ± s.d. ($n = 3$). **i**, The XCI regulatory network. Before differentiation of ESCs, *Xist* is repressed by *Tsix*-dependent and -independent mechanisms, regulated by different factors. The RNF12 protein level is low, leading to repression of *Xist* and activation of *Tsix*, respectively. Upon differentiation, the RNF12 nuclear protein concentration increases, resulting in an enhanced rate of degradation of REX1 and subsequent activation of *Xist*. **j**, Phylogenetic tree, showing the presence or absence of XCI, *Xist*, *Rnf12*, *Yy1* and *Rex1* in different species (ND, not determined; NA, not applicable).

with our combined *Xist*/*Tsix* RNA-FISH analysis and supporting a role for REX1 in activation of *Tsix* (Fig. 4f and Supplementary Fig. 9).

Genetic studies indicated that *Xist* is under control of RNF12, independent of *Tsix*[6]. In cells overexpressing REX1, upregulation of *Xist* expression during differentiation was markedly reduced (Fig. 4f and Supplementary Fig. 9). In addition, the present ChIP-seq data showed REX1 binding in the *Xist* regulatory regions, suggesting a direct action of REX1 on *Xist*. To test this in more detail, we transiently transfected male ESCs harbouring a non-functional *Tsix* stop allele with a *Rex1* expression vector. This *Rex1* overexpression resulted in suppression of *Xist*, providing further evidence for a *Tsix*-independent pathway in the repression of *Xist* by REX1 (Fig. 4g). We next performed luciferase assays with a luciferase reporter gene linked to the *Xist* promoter alone (Xist-luc), or including the distal region covering the REX1 recruitment sites in *Xist* exon1 (Xist-2p-luc). When the constructs were transiently transfected, and luciferase activity was measured at day 3 of differentiation, we found that both reporter constructs were downregulated in female *Rnf12*$^{-/-}$ and *Rex1* overexpression cell lines compared with a wild-type control female cell line (Fig. 4h). Downregulation was more prominent for the Xist-2p-luc construct in all cell lines, suggesting that REX1 represses *Xist* through the *Xist* promoter and its downstream region.

Taken together with the previous findings[9], we suggest that REX1 inhibits XCI by repression of *Xist*, and by activation of *Tsix*. This leads to a model in which, upon ESC differentiation or during development, an increased RNF12 concentration results in a decrease in the nuclear REX1 concentration. Because RNF12 is X-encoded, this effect will be more pronounced in differentiating female cells compared with male cells, allowing de-repression of *Xist* in female cells only (Fig. 4i).

Recently, *Rex1* homozygous knockout ESCs and mice have been generated[8,11]. Although no effect on XCI has been reported, *Rex1*$^{-/-}$ embryos were born at a sub-Mendelian ratio[11]. From the present study, we would expect that the threshold for initiation of XCI might be lower, both in male and female *Rex1*$^{-/-}$ embryos, than in wild-type embryos, resulting in aberrant initiation of XCI. The fact that some live *Rex1*$^{-/-}$ offspring were generated indicates that additional factors, possibly acting downstream or independently of RNF12, exert control over the XCI process. Interestingly, *Rex1* is present only in placental mammals, representing a retro-transposed copy of *Yy1* (ref. 12), a gene also implicated in the regulation of XCI[12,13] (Fig. 4j). *Rex1* is not present

in marsupials, which also lack *Xist*-mediated XCI[14]. This suggests co-evolution of *Xist* and *Rex1* in conjunction with the appearance of an XCI mechanism that requires *Xist*. In contrast to REX1, we found that YY1 expression is not upregulated in *Rnf12*$^{-/-}$ cells (Fig. 3b). Co-immunoprecipitation experiments using HEK293 cells indicate that YY1 and RNF12 interact, but YY1 is not ubiquitinated by RNF12 (Supplementary Fig. 10). Although our findings do not preclude a role for RNF12-mediated control of YY1, they clearly emphasize that the RNF12-mediated control of REX1 concerns a specific interplay, which does not occur between RNF12 and YY1 and which evolved after a retro-transposition event generated the *Rex1* retrogene. We propose that the origin of *Rex1* has played a key role in the evolution of random XCI in placental mammals.

METHODS SUMMARY

REX1 was identified as an RNF12 interaction partner by mass spectrometry analysis on Flag-affinity purified Flag–RNF12, isolated from nuclear extracts of day-3-differentiated Flag–*Rnf12* transgenic ESC lines treated with the proteasome inhibitor (MG132). Protein purification and mass spectrometry analysis were done as described in ref. 15. For the ubiquitination assay in HEK293 cells, the cells were transiently transfected with polyethylenimine (Polysciences) with the indicated expression vectors. The REX1 ChIP and ChIP-seq experiments were performed as described in ref. 16 with minor modifications. RNA-FISH was performed as described in ref. 4. For the luciferase reporter assay, ESCs were transfected with the indicated vectors, using Lipofectamine 2000 (Invitrogen). Firefly and *Renilla* luciferase activity were measured using a dual-luciferase reporter assay system (Promega).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 November 2010; accepted 9 March 2012.
Published online 29 April 2012.

1. Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190,** 372–373 (1961).
2. Jonkers, I. *et al.* RNF12 is an X-encoded dose-dependent activator of X chromosome inactivation. *Cell* **139,** 999–1011 (2009).
3. Monkhorst, K. *et al.* The probability to initiate X chromosome inactivation is determined by the X to autosomal ratio and X chromosome specific allelic properties. *PLoS ONE* **4,** e5616 (2009).
4. Monkhorst, K., Jonkers, I., Rentmeester, E., Grosveld, F. & Gribnau, J. X inactivation counting and choice is a stochastic process: evidence for involvement of an X-linked activator. *Cell* **132,** 410–421 (2008).
5. Shin, J. *et al.* Maternal Rnf12/RLIM is required for imprinted X-chromosome inactivation in mice. *Nature* **467,** 977–981 (2010).
6. Barakat, T. S. *et al.* RNF12 activates Xist and is essential for X chromosome inactivation. *PLoS Genet.* **7,** e1002001 (2011).
7. Hosler, B. A., LaRosa, G. J., Grippo, J. F. & Gudas, L. J. Expression of REX-1, a gene containing zinc finger motifs, is rapidly reduced by retinoic acid in F9 teratocarcinoma cells. *Mol. Cell. Biol.* **9,** 5623–5629 (1989).
8. Scotland, K. B., Chen, S., Sylvester, R. & Gudas, L. J. Analysis of Rex1 (zfp42) function in embryonic stem cell differentiation. *Dev. Dyn.* **238,** 1863–1877 (2009).
9. Navarro, P. *et al.* Molecular coupling of Tsix regulation and pluripotency. *Nature* **468,** 457–460 (2010).
10. Kim, J. D. *et al.* Rex1/Zfp42 as an epigenetic regulator for genomic imprinting. *Hum. Mol. Genet.* **20,** 1353–1362 (2011).
11. Masui, S. *et al.* Rex1/Zfp42 is dispensable for pluripotency in mouse ES cells. *BMC Dev. Biol.* **8,** 45 (2008).
12. Kim, J. D., Faulk, C. & Kim, J. Retroposition and evolution of the DNA-binding motifs of YY1, YY2 and REX1. *Nucleic Acids Res.* **35,** 3442–3452 (2007).
13. Donohoe, M. E. *et al.* Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. *Mol. Cell* **25,** 43–56 (2007).
14. Duret, L. *et al.* The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312,** 1653–1655 (2006).
15. van den Berg, D. L. *et al.* An Oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell* **6,** 369–381 (2010).
16. Soler, E. *et al.* A systems approach to analyze transcription factors in mammalian cells. *Methods* **53,** 151–162 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. van der Linden and C. Maduro for their help with some of the experiments, A. Inagaki, C. Andrieu-Soler, D. Warmerdam, L. Zhang and A. Mohd-Sarip for experimental advice, and C. Rougeulle and R. Jaenisch for providing the *Tsix*-stop cells. This work was supported by grants from the Netherlands Organisation for Scientific Research (NWO-TOP and NWO-VICI) and the European Research Council to J.G.

Author Contributions C.G. and J.G. designed the experiments. C.G. performed most experiments assisted by E.R. T.S.B., C.G. and J.G. generated the *Rnf12*$^{-/-}$ ES cells line. C.G., J.D., W.v.IJ. and E.M.A. performed the mass spectrometry and sequencing analysis. J.G., J.A.G. and C.G. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.G. (j.gribnau@erasmusmc.nl).

METHODS

Plasmids and antibodies. The coding sequences of *Rnf12*, *Rex1* and *Yy1* were amplified from mouse ESC complementary DNA (cDNA) and cloned into a TOPO blunt vector (Invitrogen). RNF12^{H569A,C572A} and RNF12^{del10aa} amino-acid mutants were generated by PCR site-directed mutagenesis. For mammalian expression, the wild-type and two mutant *Rnf12* coding sequences were subcloned into pCAG-Flag, a CAG-driven expression vector containing a Flag-tag (a gift from D. van den Berg) and pEGFP-N3 (Clontech) vectors; *Rex1* and *Yy1* were subcloned into pCAG-Flag–V5 and *Rex1* also into a modified pCherry-C1 vector (gifts from H. Lans). *Rex1* cDNA, *Rnf12* cDNA and truncated forms were subcloned into pGEX-6P-1 (GE Healthcare) vector for expression in bacteria. For the *Rex1* knockdown experiments, a mouse Rex1 shRNA sequence ACGGAG AGCTCGAAACTAA[9] was cloned into pSuper-GFP-Neo (Oligoengine) and a pSuper-GFP-Neo-control-shRNA was used as a control.

For the luciferase reporter constructs, DNA fragments containing the REX1 binding sites within the *Xist* promoter alone (Xist-luc, nucleotides -548 to $+47$) or including the *Xist* promoter distal region (Xist-luc-2p, nucleotides -548 to $+2161$) were amplified by PCR and cloned into the promoterless pGL4.10 [luc2] vector (Promega). All constructs were checked by DNA sequencing. Antibodies used were against V5 (Invitrogen), Flag–M2 (Sigma), NANOG (Calbiochem), OCT4 (Santa Cruz), SOX2 (R&D systems), REX1 (Abcam and Santa Cruz), SUZ12 (Upstate), RNF12 (Abnova), YY1 (Santa Cruz), ubiquitin (Enzo) and β-actin (Sigma).

Cell culture and DNA transfection. Mouse ESCs were grown and differentiated as previously described[4]. Flag–*Rnf12* and Flag–V5-*Rex1* transgenic ESC lines were generated by electroporation of *Rnf12*$^{+/-}$ (ref. 2) and wild-type female ESC lines F1 2-1 (129/Sv-Cast/Ei), with pCAG-Flag–*Rnf12* or pCAG-Flag–V5 vectors followed by puromycin selection. F1 2-1, F1 2-3 (129/Sv-Cast/Ei), 1.3 (16xms2), *Rnf12*$^{+/-}$ and *Rnf12*$^{-/-}$ ESC lines have been described[2,6]. Male 1.3 *Rnf12* overexpressing ESC lines were generated as described in Jonkers *et al.*[2]. The E14tg2a control and *Rex1*$^{+/-}$ male ESC line with a gene trap insertion in intron 3 of *Rex1* has been previously characterized[10], and was obtained from BayGenomics (gene trap clone no. XB238). ESCs were transfected using Lipofectamine 2000 (Invitrogen), according to the manufacturer's instructions. For the *Rex1* knockdown experiments, *Rnf12*$^{-/-}$ ESCs were transfected with pSuper-GFP-Neo Rex1 shRNA or control vectors and after 24 h GFP positive cells were sorted by fluorescence-activated cell sorting; 48 h later, cells were collected for RNA isolation. HEK293 cells were cultured under standard conditions in DMEM (Dulbecco's modified Eagle's medium) supplemented with 10% (v/v) FCS (fetal calf serum) and penicillin–streptomycin, and transfected with polyethylenimine (Polysciences).

Nuclear extract preparation. Unless otherwise indicated, cells were treated with proteasome inhibitor (15 μm MG132, Sigma) for 3 h before collection. We also tested the effect of the proteasome inhibitor epoxomicin (1 μM). ESCs and HEK293 cells were scraped from the culture dishes in ice-cold PBS plus protease inhibitor (Roche). Embryoid bodies grown in suspension were collected by centrifugation and washed twice in ice-cold PBS plus protease inhibitor. Nuclear extracts were prepared as described in ref. 17, but instead of being dialysed, were diluted 1:1 with buffer 0 (20 mM Hepes pH 7.6, 20% glycerol, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 0.5 mM DTT, 15 μM MG132 and protease inhibitors). To confirm phosphorylation of RNF12, female ESC nuclear extracts were incubated for 30 min at 30 °C in the presence or absence of lambda protein phosphatase (New England Biolabs).

Protein purification and mass spectrometry. Protein purification and mass spectrometry analysis were done essentially as described in ref. 15. Briefly, nuclear extract from Flag–*Rnf12* ESCs at day 3 of differentiation or Flag–V5-*Rex1* undifferentiated ESCs were incubated with Flag M2 antibody-agarose beads (Sigma) for 3 h at 4 °C, in the presence of Benzonase (Novagen). Bound proteins were eluted with Flag–tripeptide (Sigma). Elutions were pooled by trichloroacetic acid precipitation, proteins were separated by SDS–polyacrylamide gel electrophoresis and the gel was stained with a colloidal blue staining kit (Invitrogen). Mass spectrometry analysis was performed on a capillary liquid chromatography system (Nanoflow LC-MS/MS 1100 series; Agilent Technologies) coupled to a mass spectrometer (LTQ-Orbitrap; Thermo Fisher Scientific).

GST pull-down assays. Recombinant GST–*Rex1*, GST–*Rnf12* full-length cDNA and the truncated forms were expressed at 20 °C overnight in *Escherichia coli* BL21 (Invitrogen). Cells were collected and flash-frozen. Lysis buffer (50 ml; 25 mM Hepes pH 7.6, 10% glycerol, 0.5 M NaCl, 0.01% NP-40 mM, 4 mM DTT, 2.5 mM $MgCl_2$, 50 μM $ZnCl_2$, 0.15 mg ml^{-1} lysozyme and protease inhibitors) was added per litre of culture. After sonication, soluble GST fusion proteins were bound to glutathione-sepharose beads (GE Healthcare) and analysed by Coomassie staining. For *in vitro* binding assays, the GST tag was removed from the GST–REX1 fusion protein though enzymatic digestion with PreScission Protease (GE Healthcare).

RNF12-bound beads were equilibrated in buffer 100 (20 mM Hepes pH 7.6, 10% glycerol, 100 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 0.02% NP40, 0.5 mM DTT, protease inhibitors) and incubated for 2 h at 4 °C in the presence of Benzonase (Novagene) with nuclear extracts of HEK293 cells transiently expressing V5-tagged REX1 protein, or with 0.5 μg recombinant REX1 protein. Benzonase Nuclease was added to the extract to show DNA-independence of the REX1–RNF12 interaction. Bound proteins were eluted with sample buffer and analysed by immunoblotting.

Immunoprecipitation. Undifferentiated female and male ESC nuclear extracts were incubated for 2 h with REX1 antibody or control rabbit IgG (Santa Cruz), followed by addition of protein A Sepharose (Amersham) for 1 h. After washing, bound proteins were eluted with SDS sample buffer and analysed by immunoblotting with RNF12 and REX1 antibodies.

Ubiquitination assays. For the ubiquitination assay in HEK293 cells, the cells grown in 10 cm dishes were transiently transfected for 48 h with 2 μg wild-type or mutant *Rnf12* expression vectors, in the absence or presence of 2 μg V5-tagged *Rex1 or Yy1* expression vectors. Where indicated, cells were treated with proteasome inhibitor MG132 (15 μM for 3 h, Sigma) or epoxomicin (1 μM for 6 h, Sigma) before collection. Cells were collected by scraping in ice-cold PBS and nuclear extracts were prepared as described above. To detect protein expression, 10% of the nuclear extracts were used for immunoblotting with antibodies against REX1, YY1 or RNF12, and actin was used as a loading control. To recover V5-tagged REX1 and YY1, 15 μl of V5 antibody-agarose beads (Sigma) were added to the nuclear extracts and the mixture was rotated for 1.5 h at 4 °C. The beads were washed with buffer 150 (20 mM Hepes pH 7.6, 10% glycerol, 150 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 0.02% NP40, 0.5 mM DTT and protease inhibitors). Bound proteins were eluted with sample buffer and visualized by immunoblotting. Co-immunoprecipitated RNF12 was detected with RNF12 antibody, and polyubiquitinated REX1 with V5 and Ub antibodies.

The *in vitro* ubiquitination assay was done by adding recombinant REX1 (1 μg), GST–Rnf12 wild-type or mutant (0.5 μg), E1 (55 ng UBE1, Boston Biochem), E2 (300 ng UBCH5A, Boston Biochem) and His-Ub (2 μg, Sigma) to ubiquitination buffer (50 mM Tris pH 7.5, 150 mM NaCl, 5 mM $MgCl_2$, 2 mM ATP, 1 mM DTT and protease inhibitors) to a final volume of 30 μl. The reactions were incubated at 30 °C for 1 h, terminated by boiling for 5 min with sample buffer and resolved by SDS–polyacrylamide gel electrophoresis followed by immunobloting with anti-REX1 antibody.

Quantitative real-time PCR. Total RNA was extracted by using Trizol (Invitrogen) and then reverse transcribed by with Superscript II reverse transcriptase (Invitrogen) according to the manufacturer's instructions. Real-time PCR was performed using SYBR Green (Sigma) in a CFX384 real-time PCR machine (Bio-Rad). Actin was used as a normalization control. All qPCR data represent the mean ± s.d. of triplicate samples performed on cDNA isolated from three independent experiments. The primer sequences used for qPCR are listed in Supplementary Table 3; Supplementary Table 2 lists the primers used for ChIP–qPCR.

Immunofluorescence staining. ESCs were grown on coverslips without feeders and fixed with 4% paraformaldehyde for 10 min at room temperature. Subsequently, cells were permeabilized with 0.4% Triton X-100 in PBS for 10 min at room temperature and blocked with 10% goat serum in PBST (PBS with 0.05% Tween 20) for 30 min at room temperature. The coverslips were incubated with REX1 antibody overnight at 4 °C. After washing with PBST, cells were incubated with the secondary antibody, Alexa Fluor 488 goat anti-rabbit IgG (Molecular Probes) for 1 h at room temperature. After a final wash with PBS, coverslips were mounted with Vectashield Plus DAPI (Vector Laboratories). Images were acquired using a fluorescence microscope (Axioplan2; Carl Zeiss).

RNA-FISH. RNA-FISH was performed as described in (ref. 4) with minor modifications. Pre-plated female ESCs were seeded on gelatin-coated coverslips without feeders in embryoid body differentiation media (IMDM + Glutamax (GIBCO), 15% FCS, 50 μl $μl^{-1}$ ascorbic acid, NEAA, penicillin–streptomycin, 37.8 μl l^{-1} monothioglycerol (97%)). At day three of differentiation, cells were fixed and subjected to RNA-FISH. *Xist* clouds were counted for three different coverslips per cell line analysed. The *Xist* probe was a 5.5-kilobase BglII cDNA fragment covering *Xist* exons 3–7. The *Tsix* probe was a 5.1-kilobase SalI–SacII fragment of *Tsix* intron 3.

ChIP and ChIP-seq. The ChIP and ChIP-seq experiments were performed as described[16] with minor modifications. Briefly, female ESCs expressing V5-tagged REX1 and control wild-type ESCs were grown without feeders to 80% confluence (3×10^7 cells per ChIP or 1×10^8 cells per ChIP-seq). For ChIP-seq and ChIP (where indicated), the cells were treated for 3 h with proteasome inhibitor (15 μm MG132, Sigma) before chromatin was cross-linked for 10 min at room temperature with 1% formaldehyde. The cross-linking reaction was stopped by addition of 0.125 M glycine. Sonicated chromatin was immunoprecipitated with 60 μl of pre-blocked V5 antibody-agarose beads (Sigma) for each ChIP-seq.

Purified ChIP-DNA was prepared for sequencing on a HiSeq 2000 sequencer (Illumina).

The data were analysed using a combination of bioconductor packages (Shortread, ChIP-Seq and MACS). Illumina reads (36 base pairs) were aligned against the mouse genome (*Mus musculus* National Center for Biotechnology Information build 37) using Solexa Genome Analyzer ELAND software. Aligned reads were imported, filtered and normalized; coverage was calculated using Shortread and ChIP-Seq packages. The resulting coverage graph was visualized using the University of California, Santa Cruz (UCSC) Genome Browser. The significance of peaks was calculated using the MACS package. Peaks with a fold change of at least 4 and a false discovery rate of no more than 0.001 were taken as significant.

Luciferase reporter assay. ESCs were seeded into 24-well plates in differentiation medium and after 24 h transfected with 0.8 μg of the indicated vectors, using Lipofectamine 2000 (Invitrogen). To normalize for transfection efficiency, a GL4.74 (hluc/TK) vector (Promega) expressing *Renilla* luciferase was co-transfected. Firefly and *Renilla* luciferase activity were measured 48 h after transfection using a dual-luciferase reporter assay system (Promega) according to the manufacturer's instructions. Three independent experiments were performed in triplicate. The data are shown as the mean $\pm$ s.d.

17. Dignam, J. D., Lebovitz, R. M. & Roeder, R. G. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res.* **11,** 1475–1489 (1983).

doi:10.1038/nature10998

A PPARγ–FGF1 axis is required for adaptive adipose remodelling and metabolic homeostasis

Johan W. Jonker[1]†*, Jae Myoung Suh[1]*, Annette R. Atkins[1], Maryam Ahmadian[1], Pingping Li[2], Jamie Whyte[1], Mingxiao He[1], Henry Juguilon[1], Yun-Qiang Yin[1], Colin T. Phillips[1], Ruth T. Yu[1], Jerrold M. Olefsky[2], Robert R. Henry[2,3], Michael Downes[1] & Ronald M. Evans[1,4]

Although feast and famine cycles illustrate that remodelling of adipose tissue in response to fluctuations in nutrient availability is essential for maintaining metabolic homeostasis, the underlying mechanisms remain poorly understood[1,2]. Here we identify fibroblast growth factor 1 (FGF1) as a critical transducer in this process in mice, and link its regulation to the nuclear receptor PPARγ (peroxisome proliferator activated receptor γ), which is the adipocyte master regulator and the target of the thiazolidinedione class of insulin sensitizing drugs[3–5]. FGF1 is the prototype of the 22-member FGF family of proteins and has been implicated in a range of physiological processes, including development, wound healing and cardiovascular changes[6]. Surprisingly, FGF1 knockout mice display no significant phenotype under standard laboratory conditions[7–9]. We show that FGF1 is highly induced in adipose tissue in response to a high-fat diet and that mice lacking FGF1 develop an aggressive diabetic phenotype coupled to aberrant adipose expansion when challenged with a high-fat diet. Further analysis of adipose depots in FGF1-deficient mice revealed multiple histopathologies in the vasculature network, an accentuated inflammatory response, aberrant adipocyte size distribution and ectopic expression of pancreatic lipases. On withdrawal of the high-fat diet, this inflamed adipose tissue fails to properly resolve, resulting in extensive fat necrosis. In terms of mechanisms, we show that adipose induction of FGF1 in the fed state is regulated by PPARγ acting through an evolutionarily conserved promoter proximal PPAR response element within the *FGF1* gene. The discovery of a phenotype for the FGF1 knockout mouse establishes the PPARγ–FGF1 axis as critical for maintaining metabolic homeostasis and insulin sensitization.

As part of a directed screen to identify genes that respond to dietary cues in metabolic tissues (muscle, liver, brown adipose tissue (BAT) and white adipose tissue (WAT)), we observed that FGF1 is selectively induced in visceral (that is, gonadal) WAT (gWAT) in response to a high-fat diet (HFD), pointing to a possible metabolic function (Fig. 1a and Supplementary Fig. 1a, b). Subfractionation of adipose depots revealed that FGF1 was expressed in the adipocyte fraction but not in the stromal vascular fraction (SVF) of gWAT, and to a lesser extent in the adipocyte fraction of subcutaneous (that is, inguinal) WAT (iWAT) (Fig. 1b). Given that *FGF1* gene transcription is directed by at least three distinct promoters that are conserved across mammals[10] (Fig. 1c), we examined the tissue-specific expression patterns of the alternative splice variants, which differ only in their 5′ untranslated exons. Whereas the *Fgf1a* and *Fgf1b* splice variants were both expressed in gWAT (as well as other tissues) of mice fed a standard chow diet (Fig. 1d–f), only the *Fgf1a* transcript showed a striking and progressive 20-fold induction between fasted, fed and HFD exposure (Fig. 1g, h). Although no metabolic role has been attributed to FGF1, these results prompted us to reconsider this possibility. Consistent with previous reports, no metabolic or histological abnormalities, or major gene expression changes were observed in $Fgf1^{-/-}$ mice fed a standard chow diet[7] (Supplementary Figs 1c, 2; Supplementary Table 3). Similarly, when placed on an HFD, $Fgf1^{-/-}$ and wild-type cohorts

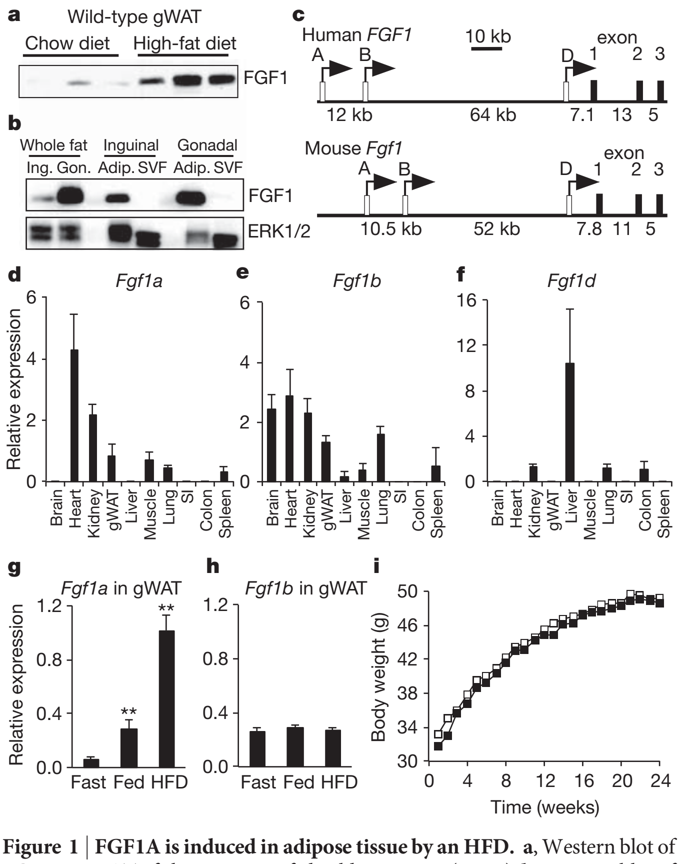


Figure 1 | FGF1A is induced in adipose tissue by an HFD. **a**, Western blot of FGF1 in gWAT of chow or HFD-fed wild-type mice ($n = 3$). **b**, Western blot of FGF1 in whole fat, adipocyte (adip.) and SVFs of inguinal (ing.) and gonadal (gon.) WAT of chow-fed wild-type mice (ERK1/2 loading control). **c**, Diagram depicting three distinct promoters driving the untranslated exons 1A, 1B and 1D (open bars) of human *FGF1* and mouse *Fgf1* genes. Alternative splicing of untranslated exons results in identical FGF1 polypeptides. **d–f**, mRNA tissue distribution in mice for *Fgf1a* (**d**), *Fgf1b* (**e**) and *Fgf1d* (**f**). **g**, **h**, mRNA levels of *Fgf1a* (**g**) and *Fgf1b* (**h**) in gWAT of overnight fasted, chow fed, or 2-weeks HFD-fed wild-type mice ($n = 5$). **i**, Body weight of HFD-fed wild-type (open symbol) and $Fgf1^{-/-}$ (filled symbol) mice over 24 weeks. Data expressed as mean ± s.d. $^{**}P < 0.01$.

[1]Gene Expression Laboratory, Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA. [2]Department of Medicine, Division of Endocrinology and Metabolism, University of California at San Diego, La Jolla, California 92093, USA. [3]Section of Diabetes/Metabolism, Veterans Affairs San Diego Healthcare System, San Diego, California 92161, USA. [4]Howard Hughes Medical Institute, Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA. †Present address: Center for Liver, Digestive and Metabolic Diseases, Department of Pediatrics, University Medical Center Groningen, University of Groningen, Hanzeplein 1, 9713 GZ Groningen, The Netherlands.
*These authors contributed equally to this work.

showed equivalent changes in weight (monitored for 24 weeks), serum adipokines, cytokines (leptin, resistin, interleukin-6 (IL-6), TNFα, tPAI-1, total and high-molecular-weight adiponectin) and serum lipids (cholesterol, free fatty acids and triglycerides) (Fig. 1i, Supplementary Tables 1 and 2). However, *Fgf1*$^{-/-}$ mice developed an exaggerated diabetic phenotype, with increased levels of fasting glucose and insulin (Fig. 2a), accompanied by severe insulin resistance (Fig. 2b, c) and markedly enhanced serum MCP1/CCL2 ($48.0 \pm 3.4\,\mathrm{pg\,ml^{-1}}$ compared to $59.0 \pm 3.4\,\mathrm{pg\,ml^{-1}}$, $P < 0.01$, in wild-type and *Fgf1*$^{-/-}$ mice, respectively), a marker of adipose macrophage infiltration and a causative factor for peripheral insulin resistance[11,12].

To further investigate the role of FGF1 in insulin sensitivity, we performed hyperinsulinaemic-euglycaemic clamp studies. The steady-state glucose infusion rate (GIR) during the clamp was about 40% lower in HFD-fed *Fgf1*$^{-/-}$ mice, reflecting decreased insulin responsiveness, and was accompanied by reductions in whole-body and insulin-stimulated glucose disposal rates (GDR and IS-GDR), indicating pronounced peripheral insulin resistance (Fig. 2d–f). The ability of insulin to suppress hepatic glucose production (HGP) was also significantly compromised in HFD-fed *Fgf1*$^{-/-}$ mice, revealing hepatic insulin resistance as well (Fig. 2g). Although HFD-fed *Fgf1*$^{-/-}$ mice had enlarged steatotic livers relative to wild-type controls (Fig. 2h, i), liver function (based on serum alanine transaminase (ALT) levels) and pancreatic function (based on islet histology and insulin secretion) appeared normal (Supplementary Fig. 3, 4). On the other hand, gWAT in *Fgf1*$^{-/-}$ mice failed to expand after HFD and exhibited pronounced structural abnormalities, as evidenced by haematoxylin and eosin staining (Fig. 2j, k, Supplementary Fig. 5). As expected with higher circulating levels of MCP1, we observed a dramatic increase in macrophage infiltration, indicating a highly inflamed gWAT (Fig. 3a). Notably, no defects were observed in iWAT (data not shown). This differential sensitivity of WAT depots to HFD stress is consistent with the known association of the visceral WAT with obesity-related pathologies including insulin resistance[13].

To explore the possibility that the metabolic dysregulation observed in *Fgf1*$^{-/-}$ mice on an HFD was associated with defects in gWAT, we performed detailed histological and molecular analyses of this tissue. Histochemistry with Masson's trichrome stain of gWAT from the HFD-fed *Fgf1*$^{-/-}$ mice revealed increased collagen deposition (blue staining) and a marked heterogeneity in adipocyte size (Fig. 3b). Quantification of adipocyte cross-sectional areas showed increases in the numbers of both small and large adipocytes in *Fgf1*$^{-/-}$ mice relative to wild-type (Fig. 3c). Next we examined the functional architecture of the adipose vasculature by intravenous injection of fluorescent microbeads. Epifluorescence microscopy of tissue sections from fluorescent-bead-perfused mice revealed decreased vascular density specifically in gWAT (Fig. 3d) but not in BAT or iWAT of HFD-fed *Fgf1*$^{-/-}$ mice (Supplementary Fig. 6). Microarray analyses identified multiple transcriptional changes induced by HFD in *Fgf1*$^{-/-}$ gWAT that were consistent with the observed phenotypes. The obesity and insulin resistance markers RBP4 and CCL11 were upregulated by

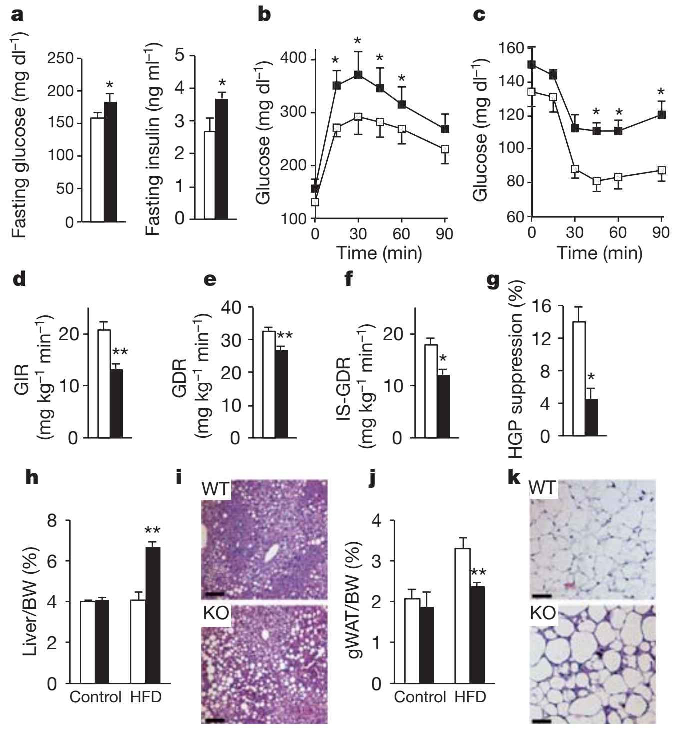


Figure 2 | Loss of FGF1 results in diet-induced insulin resistance. Metabolic studies on 24-week-old male wild-type (WT) (open bars) and *Fgf1*$^{-/-}$ (knockout, KO) (filled bars) mice fed an HFD for 16 weeks. **a**, Fasting serum glucose and insulin levels. **b**, **c**, Glucose (**b**) and insulin (**c**) tolerance tests. **d–g**, GIR (**d**), GDR (**e**), IS-GDR (**f**) and percentage suppression of HGP (**g**) during hyperinsulinaemic-euglycaemic clamp studies. **h**, Liver (percentage of body weight) from chow and HFD-fed mice ($n = 6$–7). **i**, Haematoxylin and eosin staining of liver from HFD-fed WT and KO mice. **j**, gWAT (percentage of body weight, BW) from chow and HFD-fed mice ($n = 6$–7). **k**, Haematoxylin and eosin staining of gWAT from HFD-fed WT and KO mice. Scale bar, 100 μm. Data expressed as mean ± s.d. $^{*}P < 0.05$, $^{**}P < 0.01$.

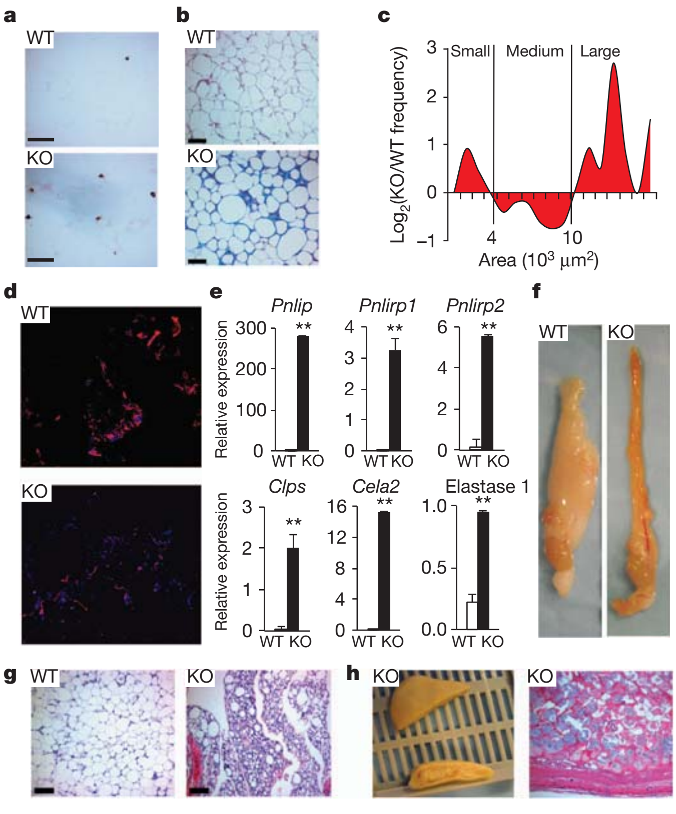


Figure 3 | Loss of FGF1 results in defects in adipose remodelling during HFD. **a**, Immunohistochemistry for the macrophage marker, F4/80 in gWAT from HFD-fed wild-type (WT) and *Fgf1*$^{-/-}$ (KO) mice. **b**, Trichrome staining for collagen deposition in gWAT from HFD-fed WT and KO mice. **c**, Quantitation of adipose cell cross-sectional area from HFD-fed mice, expressed as the ratio of the number of cells from KO to WT mice defined as small (1,000–4,000 μm^2), medium (4,000–10,000 μm^2) and large (>10,000 μm^2) cells. **d**, Fluorescence microscopy of gWAT from HFD-fed WT and KO mice perfused with microbeads (red) and counterstained with DAPI (blue). **e**, Verification by quantitative polymerase chain reaction of induction of genes associated with fat necrosis in HFD-fed WT (open bars) and KO (filled bars) mice. **f**, gWAT depots from HCC WT and KO mice. **g**, Haematoxylin and eosin staining of gWAT from HCC WT and KO mice. **h**, Image and haematoxylin and eosin staining of dissociated necrotic WAT taken from peritoneal cavity of HCC KO mice. Scale bar, 100 μm. Data expressed as mean ± s.d. $^{**}P < 0.01$.

HFD, as was the angiogenic factor VEGF. Interestingly, expression of PPARγ was also induced by HFD in $Fgf1^{-/-}$ gWAT, as was the expression of known PPARγ target genes (for example, perilipin, *Dgat1* and *Dgat2*) (Supplementary Table 3). However, the most dramatic inductions of gene expression were seen in multiple fat necrosis-associated pancreatic lipases and the tissue remodelling factor elastase 1, which were confirmed by quantitative polymerase chain reaction (qPCR; Fig. 3e, Supplementary Table 3)[14–16].

The observed histopathological and molecular changes in $Fgf1^{-/-}$ gWAT suggested a failure to execute the appropriate adipose remodelling program in response to HFD stress. As adipose tissue needs to dynamically expand and contract with fluctuations in nutrient availability, we postulated that FGF1 would also play a role in its contraction capacity on withdrawal from HFD. To test this, we re-adapted HFD-fed wild-type and $Fgf1^{-/-}$ mice to 6 weeks of chow feeding. Gross examination of HFD-to-chow converted (HCC) mice revealed features consistent with maladaptation resulting in disfigurement and discoloration of gWAT from $Fgf1^{-/-}$ mice compared to control wild-type mice (Fig. 3f). Histological examination of the HCC $Fgf1^{-/-}$ gWAT showed profound degeneration of adipose architecture and integrity of far greater severity than the HCC wild-type gWAT (Fig. 3g). Indeed, HCC $Fgf1^{-/-}$ mice frequently presented with what appeared to be fragments of dissociated fat tissue within the peritoneal cavity, which upon histological analysis were consistent with a fat necrosis pathology (Fig. 3h). The observations from the HFD and HCC regimens demonstrate a dynamic requirement for FGF1 in both adipose tissue expansion and contraction. Together, our findings show that when challenged with an HFD or HCC, $Fgf1^{-/-}$ mice are unable to remodel visceral adipose tissue in response to dietary changes. This suggests that defects in adipose plasticity, attributable to the loss of FGF1, are causally linked with a series of peripheral pathologies, including hepatic steatosis and systemic insulin resistance. These results establish FGF1 as a transducer of adipose remodelling in response to nutrient fluctuations, and identify an indispensible role for FGF1 in defending the body against metabolic disease.

As PPARγ expression was induced in $Fgf1^{-/-}$ gWAT by HFD, and HFD elevates the levels of circulating PPAR ligands[17,18], we postulated that the HFD-induction of the *Fgf1a* splice variant may be regulated by a PPAR family member. Luciferase reporter assays indeed revealed a robust induction of the human *FGF1A* transcript by the PPARs, of which PPARγ was the most potent (Fig. 4a). Furthermore, in this system the PPAR induction of FGF1 appears specific to the *FGF1A* transcript, as PPARs did not induce expression of the *FGF1B* splice variant (Fig. 4b). Examination of the promoter region of the *FGF1A* transcript revealed a highly conserved (98% conservation) region ~100 base pairs (bp) proximal to the transcription start site (Supplementary Fig. 7a) containing a conserved putative PPAR response element (PPRE) at −60 bp relative to the transcription start site (Fig. 4c). Inactivation of this PPRE using site directed mutagenesis resulted in a complete loss of its response to PPARγ (Fig. 4c, d: compare human versus ΔPPRE). To examine the functional conservation of FGF1 regulation by PPARγ, we assayed the native human and mouse *FGF1A* promoters along with reporter constructs containing orthologous PPREs (rat, horse and opossum) introduced into the human *FGF1A* promoter. PPARγ activation was observed for all promoters except for the more distantly related opossum PPRE (Fig. 4d). Chromatin immunoprecipitation (ChIP) experiments confirmed that PPARγ does indeed bind to the identified PPRE in 3T3-L1 adipocytes (Fig. 4e, Supplementary Fig. 7b). Data mining of a published genome-wide PPARγ ChIP sequencing (ChIP-Seq) study indicates that the PPARγ–*Fgf1a* promoter interaction may be specific to the adipocyte fraction within WAT, as PPARγ binding was seen in 3T3-L1 adipocytes, but not in macrophages[19]. Together, these findings show that the adipocyte PPARγ–FGF1 axis is functionally conserved in a wide range of mammals.

To confirm the physiological relevance of the PPARγ induction of FGF1, we determined the expression of the *Fgf1A* transcript in mice in

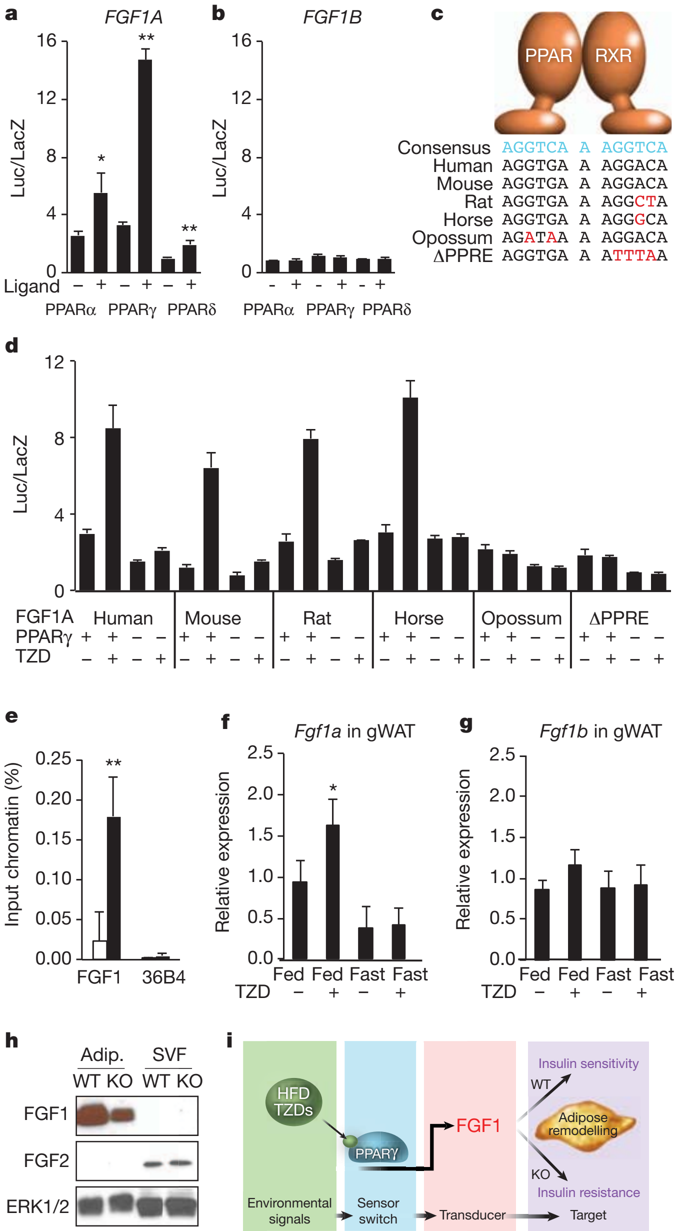


Figure 4 | *FGF1* is a direct transcriptional target of PPARγ. a, **b**, Luciferase reporter assays of *FGF1A* (**a**) and *FGF1B* (**b**) promoters co-transfected with PPARs with or without ligand. **c**, Sequence alignment of the putative PPRE, recognized by PPAR and its heterodimeric partner Retinoid X Receptor (RXR), within the *FGF1A* promoter from different species. Red indicates nucleotide variations between the PPREs relative to human. **d**, Species-specific response of the *FGF1A* promoters to PPARγ with or without ligand using luciferase reporter assays. **e**, Chromatin immunoprecipitation of PPARγ on the *Fgf1a* promoter in differentiated 3T3-L1 cells (open bars, immunoglobulin-G (IgG); filled bars, PPARγ antibody). **f**, **g**, Levels of *Fgf1a* (**f**) *Fgf1b* (**g**) mRNA in gWAT of fed or overnight fasted wild-type mice ($n = 5$) with or without rosiglitazone (TZD; 5 mg kg^{-1} for 3 days, intraperitoneally). **h**, Western blot of FGF1, FGF2 in adipocyte (adip.) and SVFs of gWAT from chow fed wild-type (WT) and *aP2-Cre; Pparg*$^{fl/fl}$ (adipocyte-specific *Pparγ* knockout) mice (KO) (ERK1/2 loading control). **i**, Model depicting the role of the PPARγ–FGF1 axis in adipose remodelling and insulin sensitivity. Data expressed as mean ± s.d. *$P < 0.05$, **$P < 0.01$.

response to the potent and specific PPARγ ligand rosiglitazone. We found that oral administration of rosiglitazone (5 mg kg^{-1} for 3 days) significantly increased the levels of the *Fgf1a* transcript in gWAT, in fed but not fasted states (Fig. 4f). The expression of *Fgf1b* and *Fgf1d* transcripts was unchanged by rosiglitazone in gWAT and liver (Fig. 4g;

Supplementary Fig. 8), whereas *Fgf1a* and *Fgf1d* transcripts were undetectable in liver and gWAT, respectively (data not shown). Furthermore, adipocyte-specific PPARγ knockout mice[20] displayed decreased levels of FGF1 in the adipocyte fraction, without compensatory changes in the closely related FGF2, which was only detected in the SVF (Fig. 4h).

We have discovered an unexpected metabolic role for FGF1 as a critical transducer of PPARγ signalling that mediates the proper coupling of nutrient storage to adaptive remodelling of adipose tissue. We found that HFD results in potent and selective induction of FGF1 in adipose tissue and that its transcription is controlled by PPARγ and its insulin sensitizing ligands. Loss of FGF1 leads to systemic metabolic dysfunction and insulin resistance, revealing an indispensable role for FGF1 in metabolic homeostasis (Fig. 4i). Importantly, we show that $Fgf1^{-/-}$ mice are unable to both properly expand and contract their adipose tissue in response to dietary changes, revealing the dynamic requirement of FGF1 for adipose tissue remodelling. The capacity of adipose tissue to remodel is crucial for accommodating changes in energy availability in fasted and fed states but is not unlimited, and can become perturbed in obesity and related pathologies. Previous reports have shown that FGF1 can signal through FGFRs to pre-adipocytes[21–23]. Recently, several endocrine FGF family members (FGF15/19, 21) have been linked to metabolic homeostasis through nuclear receptor regulation[24–26]. We now expand this nuclear receptor–FGF interface to include the paracrine FGF1. Our discovery of the PPARγ–FGF1 axis leads us to consider the therapeutic potential of FGF1 in potentially mediating insulin sensitization without provoking the full range of adverse events associated with PPARγ activation.

METHODS SUMMARY

$Fgf1^{-/-}$ mice and age- and sex-matched wild-type controls (>99% C57BL/6 genetic background) received a standard diet or high fat (60%) diet (F3282, Bio-Serv) and water *ad libitum*. For the HFD-to-chow conversion diet regimen (HCC), mice were fed HFD starting from 6 weeks of age. After 9 months of HFD, the diet was converted back to standard laboratory chow for 6 weeks. Glucose tolerance tests and insulin tolerance tests were conducted after overnight fasting and after 5 h fasting, respectively. Glucose (1 g per kg, i.p.) or insulin (0.5 U insulin per kg, i.p.) was injected and blood glucose monitored. Serum analyses were performed on blood collected by tail bleeding either in the *ad libitum* fed state or following overnight fasting. Free fatty acids, triglycerides, cholesterol and ALT were measured using commercial enzymatic colorimetric kits. Serum insulin levels and total and high-molecular weight (HMW) adiponectin levels were measured by ELISAs using commercial kits. Plasma adipokine levels were measured using a Milliplex MAP kit. Histological and immunohistochemical analyses were performed on sections of fixed tissues according to standard procedures.

Vasculature was visualized by tail vein injection of fluorescent microbeads (0.1 μm FluoSpheres, Molecular Probes). Subsequently, mice were anaesthetized and perfused through the heart with additional fluorescent microbeads. Tissues were dissected, embedded and 10-μm frozen sections analysed for blood vessel density using fluorescence microscopy.

Luciferase reporter assays were performed in CV-1 cells treated overnight with or without ligands (PPARα, 1 μM WY14643 (Cayman Chemicals); PPARγ, 1 μM rosiglitazone (TZD) (Cayman Chemicals); PPARδ, 100 nM GW1516 (Santa Cruz)). Site-directed mutagenesis of the PPRE in the human *FGF1A* promoter was performed using a QuikChange II kit.

Chromatin immunoprecipitation assays were performed on differentiated 3T3-L1 cells. Sheared chromatin generated from cross-linked cell pellets by sonication was incubated with PPARγ antibody or control rabbit IgG, and precipitated with pre-blocked protein A-agarose beads.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 22 September 2011; accepted 27 February 2012.
Published online 22 April 2012.

1. Lee, M. J., Wu, Y. & Fried, S. K. Adipose tissue remodeling in pathophysiology of obesity. *Curr. Opin. Clin. Nutr. Metab. Care* **13,** 371–376 (2010).
2. Sun, K., Kusminski, C. M. & Scherer, P. E. Adipose tissue remodeling and obesity. *J. Clin. Invest.* **121,** 2094–2101 (2011).
3. Forman, B. M. *et al.* 15-Deoxy-$\Delta^{12,14}$-prostaglandin J_2 is a ligand for the adipocyte determination factor PPARγ. *Cell* **83,** 803–812 (1995).
4. Barak, Y. *et al.* PPARγ is required for placental, cardiac, and adipose tissue development. *Mol. Cell* **4,** 585–595 (1999).
5. Tontonoz, P. & Spiegelman, B. M. Fat and beyond: the diverse biology of PPARγ. *Annu. Rev. Biochem.* **77,** 289–312 (2008).
6. Itoh, N. & Ornitz, D. M. Functional evolutionary history of the mouse *Fgf* gene family. *Dev. Dyn.* **237,** 18–27 (2008).
7. Miller, D. L., Ortega, S., Bashayan, O., Basch, R. & Basilico, C. Compensation by fibroblast growth factor 1 (FGF1) does not account for the mild phenotypic defects observed in FGF2 null mice. *Mol. Cell. Biol.* **20,** 2260–2268 (2000).
8. Beenken, A. & Mohammadi, M. The FGF family: biology, pathophysiology and therapy. *Nature Rev. Drug Discov.* **8,** 235–253 (2009).
9. Itoh, N. & Ornitz, D. M. Fibroblast growth factors: from molecular evolution to roles in development, metabolism and disease. *J. Biochem.* **149,** 121–130 (2011).
10. Myers, R. L., Payson, R. A., Chotani, M. A., Deaven, L. L. & Chiu, I. M. Gene structure and differential expression of acidic fibroblast growth factor mRNA: identification and distribution of four different transcripts. *Oncogene* **8,** 341–349 (1993).
11. Kanda, H. *et al.* MCP-1 contributes to macrophage infiltration into adipose tissue, insulin resistance, and hepatic steatosis in obesity. *J. Clin. Invest.* **116,** 1494–1505 (2006).
12. Kamei, N. *et al.* Overexpression of monocyte chemoattractant protein-1 in adipose tissues causes macrophage recruitment and insulin resistance. *J. Biol. Chem.* **281,** 26602–26614 (2006).
13. Gesta, S., Tseng, Y. H. & Kahn, C. R. Developmental origin of fat: tracking obesity to its source. *Cell* **131,** 242–256 (2007).
14. Schmitz-Moormann, P., von Wedel, R., Agricola, B. & Himmelmann, G. W. Studies of lipase-induced fat necrosis in rats. *Pathol. Res. Pract.* **163,** 93–108 (1978).
15. Lee, P. C., Nakashima, Y., Appert, H. E. & Howard, J. M. Lipase and colipase in canine pancreatic juice as etiologic factors in fat necrosis. *Surg. Gynecol. Obstet.* **148,** 39–44 (1979).
16. Chua, F. & Laurent, G. J. Neutrophil elastase: mediator of extracellular matrix destruction and accumulation. *Proc. Am. Thorac. Soc.* **3,** 424–427 (2006).
17. Barish, G. D., Narkar, V. A. & Evans, R. M. PPARδ: a dagger in the heart of the metabolic syndrome. *J. Clin. Invest.* **116,** 590–597 (2006).
18. Sugii, S. *et al.* PPARγ activation in adipocytes is sufficient for systemic insulin sensitization. *Proc. Natl Acad. Sci. USA* **106,** 22504–22509 (2009).
19. Lefterova, M. I. *et al.* Cell-specific determinants of peroxisome proliferator-activated receptor γ function in adipocytes and macrophages. *Mol. Cell. Biol.* **30,** 2078–2089 (2010).
20. He, W. *et al.* Adipose-specific peroxisome proliferator-activated receptor γ knockout causes insulin resistance in fat and liver but not in muscle. *Proc. Natl Acad. Sci. USA* **100,** 15712–15717 (2003).
21. Hutley, L. *et al.* Fibroblast growth factor 1: a key regulator of human adipogenesis. *Diabetes* **53,** 3097–3106 (2004).
22. Hutley, L. J. *et al.* A putative role for endogenous FGF-2 in FGF-1 mediated differentiation of human preadipocytes. *Mol. Cell. Endocrinol.* **339,** 165–171 (2011).
23. Fon Tacer, K. *et al.* Research resource: comprehensive expression atlas of the fibroblast growth factor system in adult mouse. *Mol. Endocrinol.* **24,** 2050–2064 (2010).
24. Moore, D. D. Physiology. Sister act. *Science* **316,** 1436–1438 (2007).
25. Kliewer, S. A. & Mangelsdorf, D. J. Fibroblast growth factor 21: from pharmacology to physiology. *Am. J. Clin. Nutr.* **91,** 254S–257S (2010).
26. Kharitonenkov, A. FGFs and metabolism. *Curr. Opin. Pharmacol.* **9,** 805–810 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Alvarez, S. Kaufman, N. H. Uhlenhaut, M. Hassan and E. Williams for technical assistance, and L. Ong and S. Ganley for administrative assistance. R.M.E. is an Investigator of the Howard Hughes Medical Institute at the Salk Institute and March of Dimes Chair in Molecular and Developmental Biology. This work was supported by National Institutes of Health grants (DK062434, DK057978, DK090962, DK063491 and HL105278), the Helmsley Charitable Trust, and the Howard Hughes Medical Institute. J.W.J. is supported by the Human Frontier Science Program (HFSP), the Netherlands Organization for Scientific Research (NWO) and an EU Marie Curie Reintegration grant (IRG-277169). M.A. is supported by an F32 Ruth L. Kirschstein National Research Service Award (NIDDK).

Author Contributions J.W.J, J.M.S, M.D. and R.M.E. designed and supervised the research. J.W.J., J.M.S., A.R.A., M.A., P.L., M.H., J.W., H.J., Y.-Q.Y. and C.T.P. performed research. R.R.H. provided samples and analysed results. J.W.J., J.M.S., R.T.Y., J.M.O., M.D. and R.M.E. analysed data. J.W.J, J.M.S., A.R.A., M.A., M.D. and R.M.E. wrote the manuscript.

Author Information Microarray data sets have been deposited in the NCBI Gene Expression Omnibus, accession number GSE31692. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.M.E. (evans@salk.edu) or M.D. (downes@salk.edu).

METHODS

Animals. *Fgf1*$^{-/-}$ mice[7] and age- and sex-matched wild-type controls (>99% C57BL/6 genetic background) received a standard chow diet (MI laboratory rodent diet 5001, Harlan Teklad) or high fat (60%) diet (F3282, Bio-Serv) and water *ad libitum*. *Pparg*$^{fl/fl}$ mice were crossed with *aP2-Cre* mice to generate *aP2-Cre; Pparg*$^{fl/fl}$ mutant mice as previously described[20] and received standard chow up to analysis at 5 months of age. All mice used for studies were male unless otherwise noted.

Reporter assays. Luciferase reporter assays were performed in CV-1 cells treated overnight with or without ligands (PPARα, 1 μM WY14643; PPARγ, 1 μM Rosiglitazone (TZD); PPARδ, 100 nM GW1516). Site-directed mutagenesis of the PPRE in the human *FGF1A* promoter was performed using a QuikChange II kit (Stratagene).

Western analysis. Total cell lysates prepared in 50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1% Triton-X100, 0.1% SDS, 2 mM sodium azide and protease inhibitor cocktail (Complete, Roche), were resolved by SDS–PAGE and probed using primary antibodies to FGF1, FGF2 (Santa Cruz) and ERK1/2 (Cell Signaling Technology).

Serum analysis. Blood was collected by tail bleeding either in the *ad libitum* fed state or following overnight fasting. Free fatty acids (Wako), triglycerides (Thermo), cholesterol (Thermo) and ALT (Thermo) were measured using enzymatic colorimetric methods. Serum insulin levels (Ultra Sensitive Insulin, Crystal Chem) and total and high-molecular weight (HMW) adiponectin levels (ALPCO) were measured by ELISAs. Plasma adipokine levels were measured using a Milliplex MAP kit (Millipore).

Histological analysis and immunohistochemistry. Sections (4 μm) of fixed tissues were stained with haematoxylin and eosin according to standard procedures. For immunohistochemistry, tissues were deparaffinized in xylene and rehydrated. Slides were incubated with 5% normal donkey serum for 30 min, followed by overnight incubation with primary and secondary antibodies (F4/80, Abcam, 1:100).

Adipocyte size analysis. Adipocyte cross-sectional area was determined from photomicrographs of gonadal, mesenteric and inguinal fat pads using ImageJ[27].

Adipose tissue fractionation. Adipose tissues were excised and finely minced with a razor blade. Minced tissue was digested in adipocyte isolation buffer (100 mM HEPES pH 7.4, 120 mM NaCl, 50 mM KCl, 5 mM glucose, 1 mM $CaCl_2$, 1.5% BSA) containing 1 mg ml^{-1} collagenase at 37 °C with constant slow shaking (~120 r.p.m.) for 2 h. During the digestion period, the suspension was gently mixed several times. The suspension was then passed through a 200-μm mesh and a 100-μm mesh successively. The flowthrough was allowed to stand for 15 min to separate the floating adipocyte fraction and infranatant containing the stromal vascular fraction. The infranatant was removed and saved while minimally disturbing the floating adipocyte fraction. Both fractions were centrifuged at 500*g* for 10 min and further washed twice in DMEM/Ham's F-12 media before further manipulation.

Metabolic studies. Glucose tolerance tests and insulin tolerance tests were conducted after overnight and 5 h fasting, respectively[28,29]. Glucose (1 g kg^{-1}, i.p.) or insulin (0.5 U insulin kg^{-1}, i.p.) was injected and blood glucose monitored using a OneTouch Ultra glucometer (Lifescan).

Hyperinsulinaemic-euglycaemic clamp. Mouse clamps were performed as previously described[27,28]. Briefly, mice implanted with dual jugular catheters 3 days prior were fasted for 6 h, then equilibrated with tracer (5.0 μCi h^{-1}, 0.12 ml h^{-1} [3-^{3}H]D-glucose, NEN Life Science Products) for 90 min. A basal blood sample was then drawn via tail vein to calculate basal glucose uptake. The insulin (8 mU kg^{-1} min^{-1} at 2 μl min^{-1}, Novo Nordisk) plus tracer (5.0 μCi h^{-1}) and glucose (50% dextrose at variable rate, Abbott) infusions were initiated simultaneously, with the glucose flow rate adjusted to reach a steady state blood glucose concentration (~120 min). Steady state was confirmed by stable plasma tracer counts during the final 30 min of clamp. Blood was taken at 110 and 120 min for the determination of tracer-specific activity. At steady state, the rate of glucose disappearance or the total GDR is equal to the sum of the rate of endogenous or HGP plus the exogenous GIR. The IS-GDR is equal to the total GDR minus the basal glucose turnover rate.

Gene expression analysis. Total RNA was isolated from mouse tissue and cells using TRIzol reagent (Invitrogen). Complementary DNA was synthesized from 1 μg of DNase-treated total RNA using SuperScript II reverse transcriptase (Invitrogen). mRNA levels were quantified by qPCR with SYBR Green (Invitrogen). Samples were run in technical triplicates and relative mRNA levels were calculated by using the standard curve methodology and normalized against *36B4* mRNA levels in the same samples.

Microarrays. Total RNA (500 ng) extracted from gWAT (24-week-old wild-type and *Fgf1*$^{-/-}$ mice on chow or HFD) using TRIzol reagent (Invitrogen) was reverse transcribed into cRNA and biotin-UTP labelled using the Illumina TotalPrep RNA Amplification Kit (Ambion). cRNA was hybridized to the Illumina MouseRef-8 v2.0 Expression BeadChip using standard protocols (Illumina). Image data was converted into unnormalized sample probe profiles using the Illumina BeadStudio software and analysed by VAMPIRE. Stable variance models were constructed for each experimental conditions ($n = 2$). Differentially expressed probes were identified (unpaired VAMPIRE significance test with a two-sided, Bonferroni-corrected threshold of $\alpha_{Bonf} = 0.05$) and the significance of apparent differences between two experimental conditions determined. Lists of altered genes were mapped to pathways (VAMPIRE tool Goby) to determine KEGG categories overrepresented (Bonferroni error threshold of $\alpha_{Bonf} = 0.05$).

Chromatin immunoprecipitation assay (ChIP). Adipocytes differentiated from 3T3-L1 cells[29] (ATCC) were sequentially cross-linked with 2 mM disuccinimidyl glutarate for 30 min at room temperature (22–24 °C) and 1% formaldehyde for 10 min at room temperature. Crosslinking was quenched with glycine, and the washed cell pellet frozen at −80 °C. Cell pellets were lysed and centrifuged at 12,000*g* for 1 min at 4 °C. Sheared chromatin generated from cell pellets by sonication was incubated with PPARγ antibody (2 μg sc-7196, overnight at 4 °C) or control rabbit IgG (sc-2027, Santa Cruz Biotechnologies). The immuno-complexes were precipitated with 20 μl pre-blocked protein A-agarose beads (1 h at 4 °C) and washed extensively[30].

Input DNA isolation, DNA de-crosslinking, purification and analysis. Sheared input chromatin was ethanol precipitated and Chelex 100 (100 μl of 10% slurry, Bio-Rad) added to both input DNA and washed ChIP samples. Samples were vortexed, boiled for 10 min and then centrifuged (12,000*g* for 1 min). Samples were treated with proteinase K (1 μl of 20 mg ml^{-1}, 55 °C for 30 min), heat deactivated, and the DNA purified (Qiagen MinElute PCR Purification Kit) before qPCR analysis using the following primers. FGF1 fwd 5′-AGAGTAG GGCACAGACACAGC-3′; FGF1 rev. 5′-TGGATTAGACACGCAGGCTA-3′. aP2 fwd 5′-ATTTGCCTTCTTACTGGATCAGAGTT-3′; aP2 rev. 5′-TTGGG CTGTGACACTTCCAC-3′. Angpl4 fwd 5′-CCAGCCAGGGAAAGTAGGAG A-3′; Angpl4 rev. 5′-CAGAAAGTGCCTGCATGCC-3′. 36b4 fwd 5′-GCCA ATAGACGCGCATGTTT-3′; 36b4 rev. 5′-TGGTTCCATCGACTGTCCTG-3′.

Fluorescent microbead perfusion for vasculature studies. Mice were tail vein injected with 150 μl of PBS fluorescent microbeads (0.1 μm red fluorescent microbeads, Invitrogen), anaesthetized 5 min later, then perfused through the heart with 6 ml of 1:10 PBS dilution of fluorescent microbeads. Tissues were then dissected and embedded in Tissue-TekOCT compound (Sakura) and 10-μm frozen sections were mounted in Vectashield medium (Vector Laboratories) for analysis of blood vessel density using fluorescence microscopy[31].

Statistical analysis. All values are given as means ± standard errors. The two-tailed unpaired Student's *t*-test was used to assess the significance of difference between two sets of data. Differences were considered to be statistically significant when $P < 0.05$.

27. Fang, S. *et al.* Corepressor SMRT promotes oxidative phosphorylation in adipose tissue and protects against diet-induced obesity and insulin resistance. *Proc. Natl Acad. Sci. USA* **108,** 3412–3417 (2011).
28. Hevener, A. L. *et al.* Muscle-specific *Pparg* deletion causes insulin resistance. *Nature Med.* **9,** 1491–1497 (2003).
29. Nofsinger, R. R. *et al.* SMRT repression of nuclear receptors controls the adipogenic set point and metabolic homeostasis. *Proc. Natl Acad. Sci. USA* **105,** 20021–20026 (2008).
30. Barish, G. D. *et al.* Bcl-6 and NF-κB cistromes mediate opposing regulation of the innate immune response. *Genes Dev.* **24,** 2760–2765 (2010).
31. Springer, M. L., Ip, T. K. & Blau, H. M. Angiogenesis monitored by perfusion with a space-filling microbead suspension. *Mol. Ther.* **1,** 82–87 (2000).

doi:10.1038/nature11085

Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic

Aaron A. Thompson[1]*, Wei Liu[1]*, Eugene Chun[1]*, Vsevolod Katritch[1], Huixian Wu[1], Eyal Vardy[2], Xi-Ping Huang[2], Claudio Trapella[3], Remo Guerrini[3], Girolamo Calo[4], Bryan L. Roth[2], Vadim Cherezov[1] & Raymond C. Stevens[1]

Members of the opioid receptor family of G-protein-coupled receptors (GPCRs) are found throughout the peripheral and central nervous system, where they have key roles in nociception and analgesia. Unlike the 'classical' opioid receptors, δ, κ and μ (δ-OR, κ-OR and μ-OR), which were delineated by pharmacological criteria in the 1970s and 1980s, the nociceptin/orphanin FQ (N/OFQ) peptide receptor (NOP, also known as ORL-1) was discovered relatively recently by molecular cloning and characterization of an orphan GPCR[1]. Although it shares high sequence similarity with classical opioid GPCR subtypes (~60%), NOP has a markedly distinct pharmacology, featuring activation by the endogenous peptide N/OFQ, and unique selectivity for exogenous ligands[2,3]. Here we report the crystal structure of human NOP, solved in complex with the peptide mimetic antagonist compound-24 (C-24) (ref. 4), revealing atomic details of ligand–receptor recognition and selectivity. Compound-24 mimics the first four amino-terminal residues of the NOP-selective peptide antagonist UFP-101, a close derivative of N/OFQ, and provides important clues to the binding of these peptides. The X-ray structure also shows substantial conformational differences in the pocket regions between NOP and the classical opioid receptors κ (ref. 5) and μ (ref. 6), and these are probably due to a small number of residues that vary between these receptors. The NOP–compound-24 structure explains the divergent selectivity profile of NOP and provides a new structural template for the design of NOP ligands.

The pharmacological effects of NOP are complex and distinct from classical opioid receptors. N/OFQ shares sequence similarity with other opioid peptides, notably the κ-OR endogenous ligand dynorphin A, but does not interact with δ-OR, κ-OR or μ-OR. Similarly, the classical opioid peptides have very low affinity for NOP. Unlike the classical opioid receptors, NOP is also insensitive to most morphine-like small molecules including naloxone, thereby yielding a pharmacologically important discriminatory feature between NOP and classical opioid receptors. Studies with N/OFQ, NOP-selective agonists or antagonists, and receptor- or peptide-deficient mice have shown that the NOP system has important roles in the control of central and peripheral functions including pain, anxiety and mood, food intake, learning and memory, locomotion, cough and micturition reflexes, cardiovascular homeostasis, intestinal motility and immune responses[7]. Understanding the structural requirements for NOP ligand selectivity and modes of binding is therefore paramount for the optimization of future agonist- and antagonist-based therapeutics.

The 3.0 Å resolution X-ray crystal structure of the human NOP receptor in complex with C-24 was determined by replacing 43 N-terminal residues of the receptor with thermostabilized apocytochrome b_{562}RIL (BRIL)[8], and by truncating 31 carboxy-terminal residues of NOP (Fig. 1a) (see Methods). We found that this BRIL–NOP fusion is functional and responds to N/OFQ and the small molecular agonist SCH-221510 (ref. 9), activating endogenous heterotrimeric G_i/G_o proteins in HEK293T cells, albeit with reduced potency and efficacy (Supplementary Tables 2 and 3), perhaps owing to the C-terminal NOP truncation. C-24 was selected for co-crystallization on the basis of the pronounced thermostability it imparts on the receptor (Supplementary Fig. 1), its high affinity (half-maximum inhibitory concentration (IC_{50}) = 0.27 nM) and antagonist potency (IC_{50} = 0.1 nM) for NOP, and its selectivity (≥1,000-fold)[4]. Peripherally administered C-24 is able to penetrate the central nervous system, where it antagonizes N/OFQ effects on nociception[10] and produces beneficial responses in experimental models of Parkinson's disease[11]. The NOP structure revealed C-24 binding deep within the orthosteric binding pocket (Fig. 1a), probably mimicking the 'message' domain of N/OFQ (Phe 1-Gly 2-Gly 3-Phe 4), a similar sequence to that of canonical opioid peptides (Tyr 1-Gly 2-Gly 3-Phe 4)[7,12] (Supplementary Fig. 2).

Structural comparison of published GPCR crystal structures shows a modularity of the seven-transmembrane helical core, and considerable variation of the extracellular module with boundaries defined by proline-induced kinks[13]. NOP contains five such kinks in the seven-transmembrane core located at residue positions Pro $105^{2.58}$, Pro $184^{4.59}$, Pro $227^{5.50}$, Pro $278^{6.50}$ and Pro $316^{7.50}$ (superscripts indicate residue numbers as per the Ballesteros–Weinstein nomenclature[14]), yielding repercussions on the shape of the ligand-binding pocket. Notably, the extracellular tip of helix V in NOP is shifted by more than 4 Å as compared with the κ-OR[5] and μ-OR[6] crystal structures (Protein Data Bank (PDB) accessions 4DJH and 4DKL, respectively), thereby resulting in both a gap between helices IV and V (~12 Å between Cα of residues 184 and 215) and an expansion of the orthosteric pocket (Supplementary Fig. 5). However, compared with the chemokine receptor CXCR4, the extracellular tip of helices VI and VII are tilted inwards, towards the orthosteric pocket. Unlike the κ-OR structure[5], the extracellular half of helix I in NOP is pulled in towards the axis of the seven-transmembrane bundle, in a conformation that is more similar to that of the chemokine receptor (PDB accession 3ODU (ref. 15); Fig. 1b). This alternative conformation of helix I is facilitated by the presence of flexible glycine residues located at an apparent 'hinge point' that are conserved within the opioid receptor family: Gly $65^{1.46}$ and Gly $68^{1.49}$ in NOP, and Gly $73^{1.46}$ and Gly $76^{1.49}$ in κ-OR. NOP has an extra glycine at the hinge point, Gly $64^{1.45}$, adding to the potential flexibility of this helix.

Despite low sequence conservation, extracellular loops (ECLs) 1 and 2 of NOP are structurally similar to those of κ-OR and CXCR4 (Fig. 1b). Specifically, the backbone of ECL1 in NOP is nearly indistinguishable from that of κ-OR and CXCR4. ECL2 forms an elongated β-hairpin, which is tethered to the extracellular tip of helix III by a structurally conserved disulphide bond between Cys $123^{3.25}$ and Cys 200^{ECL2}. This β-hairpin motif is also observed in κ-OR and CXCR4, suggesting a common structural motif of the γ-branch[16] class A peptide-binding

[1]Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037, USA. [2]National Institute of Mental Health Psychoactive Drug Screening Program, Department of Pharmacology and Division of Chemical Biology and Medicinal Chemistry, University of North Carolina Chapel Hill Medical School, Chapel Hill, North Carolina 27599, USA. [3]Department of Pharmaceutical Sciences and LTTA (Laboratorio per le Tecnologie delle Terapie Avanzate), University of Ferrara, 44121 Ferrara, Italy. [4]Department of Experimental and Clinical Medicine, Section of Pharmacology and National Institute of Neuroscience, University of Ferrara, 44121 Ferrara, Italy.
*These authors contributed equally to this work.

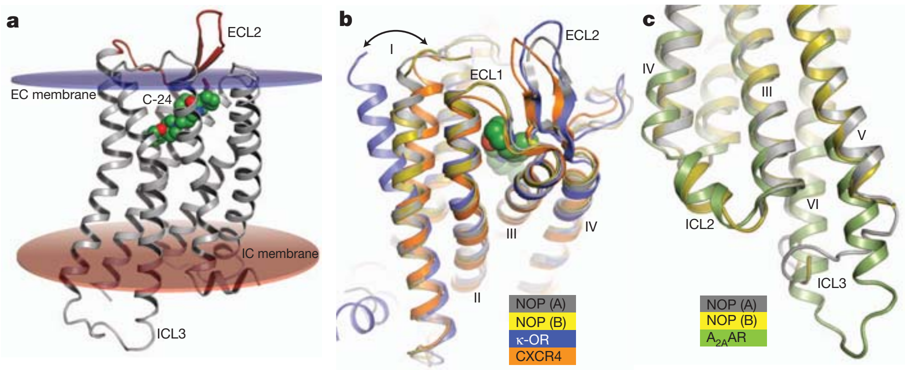


Figure 1 | Structural overview of the NOP receptor. **a**, NOP (grey) is shown in ribbon representation with its ECL2 highlighted (red). The bound ligand C-24 is depicted as green spheres, and transparent disks highlight the extracellular (EC) and intracellular (IC) membrane boundaries (coloured blue and red, respectively). **b**, Structural superposition of NOP molecules 'A' and 'B', κ-OR (PDB accession 4DJH)[5] and CXCR4 (PDB accession 3ODU)[15], coloured grey, yellow, blue and orange, respectively. Compared with κ-OR, the extracellular portion of helix I from NOP is tilted inwards towards the orthosteric pocket, in a similar conformation to CXCR4. **c**, Structural superposition of NOP molecules A and B and thermostabilized A_{2A}AR (PDB accession 3PWH)[18], coloured grey, yellow and green, respectively, highlighting conformational differences between the ICLs.

receptors. Unlike δ-OR and μ-OR, the ECL2s of κ-OR and NOP are enriched in aspartate and glutamate residues, making the loops and the entrance to their binding pocket very acidic (Supplementary Fig. 6). Moreover, ECL2 in NOP is two residues shorter than in κ-OR, making it and differences in charge distribution possible determinants for selectivity. These details are consistent with N/OFQ–dynorphin A chimaera peptide data showing that replacement of six residues on the address domain of N/OFQ with the corresponding residues from dynorphin A markedly impaired affinity and activity towards NOP[17].

Intracellular loop (ICL) 2 of NOP receptor molecule 'B' in the asymmetric unit forms a short α-helix, which has been observed in many other GPCRs (Fig. 1c); the ICL2 is tethered to the seven-transmembrane core via a salt bridge between Arg 162^{ICL2} found in all opioid receptors and Asp $147^{3.49}$ from the conserved 'D(E)RY' motif (Supplementary Fig. 7). The ICL3 connecting helices V and VI is thought to be highly malleable as it accommodates activation-related rearrangements in these helices and binding of heterotrimeric G_i/G_o proteins. Therefore, the structure of ICL3 found in the NOP receptor molecule 'A', which has 15 residues in a coil conformation and a hydrogen bond between the backbone carbonyl of Val 245^{ICL3} and the Arg $259^{6.31}$ side chain, is probably just one of the possible configurations the loop can adopt. Structural alignment with thermostabilized A_{2A}AR (PDB accession 3PWH)[18] also suggests dissimilarity in the ICL3 regions of these receptors, where NOP helices V and VI are shorter and further apart than in A_{2A}AR, and the coiled part of ICL3 is longer than A_{2A}AR (15 residues in NOP versus 8 residues in A_{2A}AR) (Fig. 1c and Supplementary Fig. 7).

The NOP–C-24 structure highlights specific residues in the pocket that are essential for N/OFQ binding and receptor subtype selectivity (Fig. 2). The orthosteric binding pocket of NOP is relatively large, reflecting its ability to bind large endogenous peptides. With a similar pose in both NOP molecules (root mean squared deviation (r.m.s.d.) = 0.6 Å), C-24 interacts with the 'floor' of the pocket through several hydrophobic and electrostatic interactions. Mutagenesis of the binding pocket of NOP defined the relative impact of specific residues on C-24 and N/OFQ binding and function (Supplementary Tables 4 and 5). The protonated nitrogen of the C-24 piperidine ring forms a crucial salt bridge with Asp $130^{3.32}$ — a residue that is conserved in the opioid receptor family and all biogenic amine GPCRs. Mutations of Asp $130^{3.32}$ to either alanine or asparagine abrogate N/OFQ binding, highlighting the requirement of the negative charge at this position[19] (Fig. 2 and Supplementary Tables 4 and 5), and it has been proposed that Asp $130^{3.32}$ is involved in a salt-bridge interaction with the positively charged N-terminal nitrogen of N/OFQ[20,21]. In addition to the anchoring salt bridge between Asp $130^{3.32}$ and the amino moiety of C-24, the linked benzofuran/piperidine rings are buried in a hydrophobic pocket created by residues from helices III, V and VI. The benzofuran 'head' group is sandwiched between Met $134^{3.36}$ and Tyr $131^{3.33}$, in which the $Met^{3.36}$ side chain adopts a different, more buried rotamer as compared to κ-OR, thereby allowing a deeper penetration of the C-24 ring system. This is consistent with the modest effect of a Met $134^{3.36}$Ala mutation on the potency of NOP ligands (Supplementary Tables 4 and 5). A Tyr $131^{3.33}$Phe mutation had no effect on agonist binding, whereas Tyr $131^{3.33}$Ala was deleterious (Supplementary Tables 4 and 5), suggesting that Tyr 131 participates in π-stacking interactions with Phe 1 of the peptide[19].

At the 'tail' end of C-24, the carbonyl group adjacent to the pyrrolidine ring is hydrogen bonded to Gln $107^{2.60}$, a residue stabilized by a hydrogen bond to Tyr $309^{7.43}$. A Gln $107^{2.60}$Ala mutation results in a 10-fold loss in C-24 binding and a more than 300-fold reduction in N/OFQ potency, and mutation of Tyr $309^{7.43}$ abolishes binding of C-24 and reduces N/OFQ potency ~7-fold (Supplementary Tables 4 and 5). Interestingly, both $Gln^{2.60}$ and $Tyr^{7.43}$ are present in the κ-OR structure, albeit in very different conformations (Supplementary Fig. 8).

The crystal structure of NOP in complex with C-24 afforded us a unique opportunity to determine the molecular basis for both the high-affinity binding by N/OFQ-derived peptide antagonists and their pronounced subtype selectivities (Fig. 3). Notably, we verified that the C-24 binding mode can be reliably reproduced by energy-based docking of C-24 to the NOP receptor, with an r.m.s.d. of ~0.9 Å. Moreover, docking of another piperidine derivative, compound-35 (C-35)[22], closely mimics the binding of C-24, whereas docking of a less active stereoisomer compound-36 (ref. 22) yields a considerably distorted binding pose in the pyrrolidine region and a reduced binding score (not shown). C-24 has previously been proposed[22] to mimic the N-terminal four residues of N/OFQ-related peptide antagonists [Nphe 1]N/OFQ(1-13)-NH_2 (in which Nphe denotes *N*-benzylglycine)[23] and UFP-101 (ref. 24). Automated docking of the four N-terminal residues of UFP-101 results in a conformation of the Nphe 1-Gly 2-Gly 3-Phe 4 tetrapeptide in which the Nphe 1 and Phe 4 rings of the peptide make the same hydrophobic interactions as the aromatic rings of C-24, and the N-terminal amino group forms a salt bridge with Asp $130^{3.32}$, thus supporting the proposed similarity in the binding poses between small molecules and peptide analogues (Fig. 3c).

The 'address' domain of N/OFQ (residues 5–17) was previously shown by NMR to have a strong preference for α-helical secondary structure[25,26], which is probably preserved in UFP-101 as the only difference in this domain are the mutations Leu14Arg and Ala15Lys. Docking of the full-length UFP-101 suggests a plausible fit of the α-helical address domain into the binding pocket entrance shaped

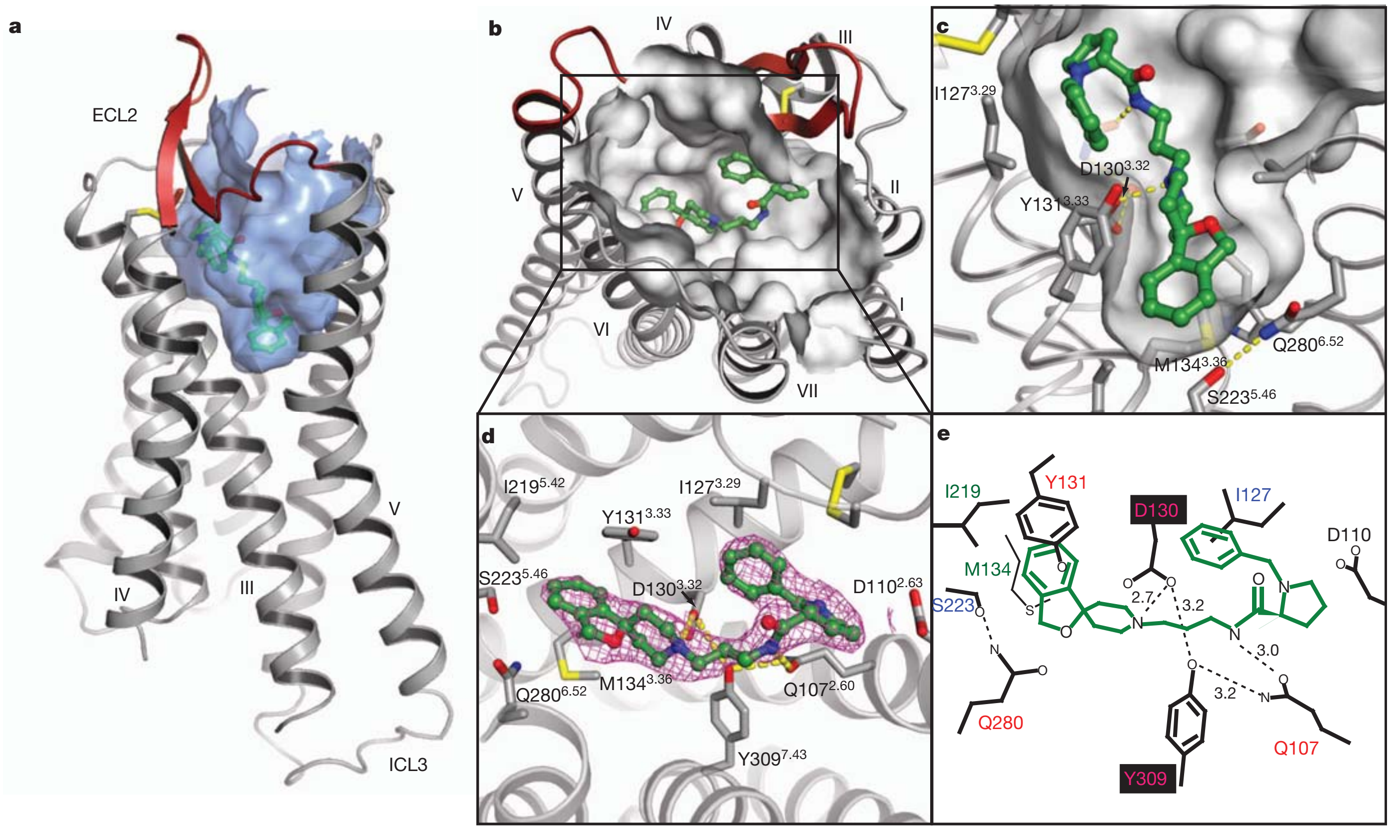


Figure 2 | The orthosteric ligand-binding pocket. **a**, Cartoon representation of NOP with its large orthosteric ligand-binding pocket shown as a blue transparent surface. ECL2 is coloured red in all subsequent figures. **b**, Extracellular view of the pocket with bound C-24 depicted as green sticks. **c**, Side view of C-24 in the binding pocket with yellow dashed lines highlighting hydrogen bond interactions and salt bridges. **d**, A σA-weighted $2mF_o - DF_c$ electron density map contoured at 1.0σ ($0.0173\,e\,Å^{-3}$) around C-24 inside the ligand-binding pocket. **e**, Schematic representation of C-24 interactions with NOP (B), with labelled distances (Å). Residue labels are coloured according to the effect on C-24 binding when replaced with alanine. Magenta-labelled residues on black background abolish C-24 binding; red-labelled residues result in an approximately 10-fold decrease in affinity; green-labelled residues slightly increase the affinity of C-24; blue-labelled residues were not tested. Asp 110 had no effect on the binding of C-24, although it is crucial for N/OFQ binding.

by the highly acidic tip of ECL2 and helices II and VII, with all six basic residues of the peptide forming ionic interactions with acidic side chains of NOP (Fig. 3c–e).

Interactions of the address domain of N/OFQ(1–13) with helices II (residues 107–113)[27] and VII (residues 296–302)[19] were previously demonstrated by photocrosslinking, a finding consistent with our mutagenesis data showing the crucial importance of Asp 110$^{2.63}$ in the binding of N/OFQ but not small molecule antagonists or the agonist SCH-221510 (Supplementary Table 5). These results suggest a similar binding mode for the address domains of N/OFQ-derived peptides. On the other hand, note that the κ-OR-binding peptide dynorphin A has a Pro 10 in the middle of the address sequence[12] that is unfavourable for α-helix formation, suggesting potential differences in the binding mode for this classical opioid peptide.

As mentioned earlier, NOP displays markedly reduced affinities for morphine-based small molecules and the classical opioid receptor peptide ligands: N/OFQ contains an N-terminal FGGF instead of the YGGF motif found in the classical opioid receptor peptide ligands. Previous biochemical studies attributed this distinct selectivity profile to the three residue positions in the binding pocket of NOP that differ from all other opioid receptors: Ala 216$^{5.39}$ (Lys in others), Gln 280$^{6.52}$ (His in others) and Thr 305$^{7.39}$ (Ile in others). Mutation of these three positions on the N/OFQ receptor to classical opioid receptor residues has been shown to be sufficient for conferring high-affinity binding to a dynorphin-derived κ-OR selective peptide[28,29]. Moreover, the same three mutations conferred nanomolar-range NOP binding of morphine-based opioid antagonists such as bremazocine, naltrexone and naltrindole, as well as a κ-OR specific antagonist norbinaltorphimine (nor-BNI)[29]. The crystal structures of NOP and κ-OR show that the side chains of these three residues are pointing towards the interior of the binding pocket (Fig. 4 and Supplementary Fig. 8). In NOP, Gln 280$^{6.52}$ and Thr 305$^{7.39}$ are involved in C-24 interactions, and all three of the cognate residues at these positions are involved in κ-OR interactions with the selective antagonist JDTic and with the modelled nor-BNI antagonists[5]. Notably, although most of the modified side chains are polar, none form direct hydrogen bonding interactions to the ligands tested, so that the selectivity profiles cannot be explained by simple polar-to-hydrophobic (or vice versa) changes of ligand contacts. Instead, a comparison of the NOP and κ-OR structures shows that several of the NOP-specific side-chain changes, including two of the substitutions mentioned earlier (Ala$^{5.39}$Lys and Gln$^{6.52}$His), are involved in a large-scale reshaping of the binding pocket and an alternative coordination of water molecules (Fig. 4). Located closer to the ligand-binding pocket entrance, Lys 227$^{5.39}$ in κ-OR, and potentially in other classical opioid receptors, is involved in salt bridges with the side chains of Asp 223$^{5.35}$ and Glu 297$^{6.58}$ (Fig. 4a). Replacement of Lys$^{5.39}$ to alanine in NOP precludes these stabilizing ionic interactions and is accompanied by an outwards shift of the extracellular half of helix V in the NOP crystal structure, and an inwards shift of helix VI. Opioid receptor subtype alteration of the large Lys$^{5.39}$ side chain and the accompanying shifts of the α-helices reshape the entrance to the pocket, and this probably affects the binding of address domains of peptides and synthetic ligands.

The κ-OR structure reveals a cluster of water molecules that is coordinated by two of the classical opioid receptor-specific residues involved in binding pocket remodelling (Fig. 4b) — His 291$^{6.52}$ and the

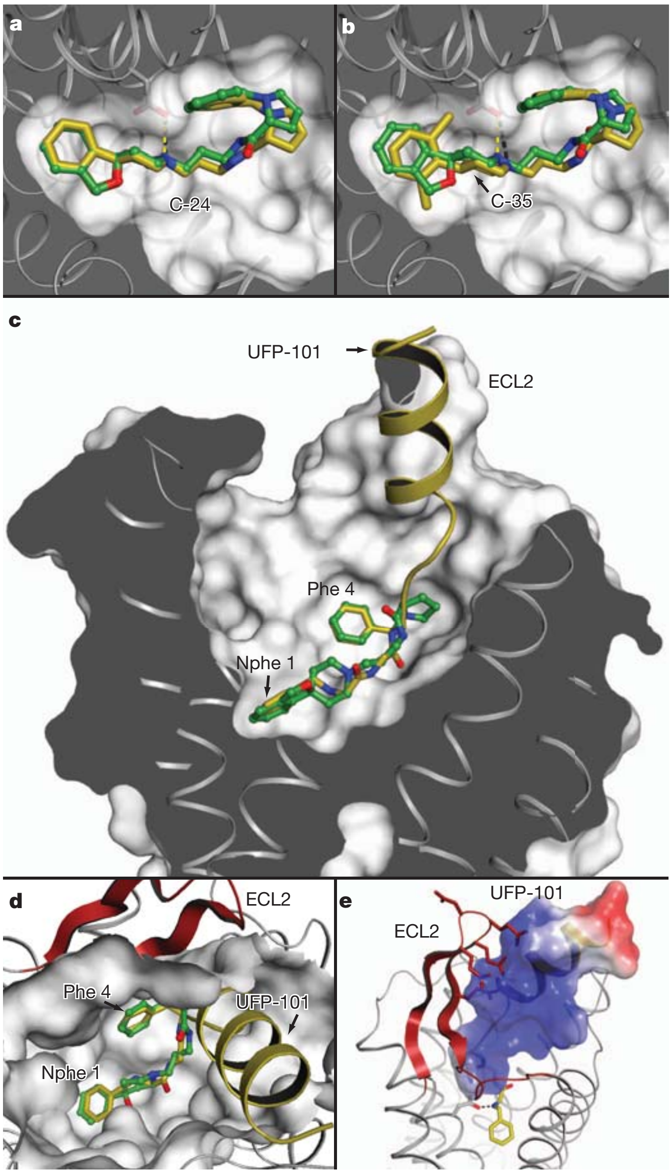


Figure 3 | Molecular docking in the orthosteric-binding pocket. a–e, The docking of C-24 (**a**), its analogue C-35 (**b**) and peptide antagonist UFP-101 (**c–e**) in the NOP. The crystallographic pose of C-24 is green in all panels, and the docked molecules (C-24, C-35 and UFP-101) are coloured yellow. The Nphe 1-Gly 2-Gly 3-Phe 4 tetrapeptide portion of the docked UFP-101 is depicted as sticks, and the 'address' domain (residues 5–17) of this peptide is represented as a cartoon. A 'sliced' side-view of the pocket is shown in **c**, and a view from the extracellular surface is shown in **d**. **e**, The electrostatic surface potentials of the UFP-101 peptide, coloured blue to red, corresponding to positive and negative surface potentials (+3 to $-3\,\mathrm{kT\,e^{-1}}$), respectively. ECL2 is coloured red, and the acidic Asp and Glu residues from the ECL2 β-hairpin are depicted as red sticks.

backbone carbonyl of Lys 227$^{5.39}$. Interestingly, one such tightly bound water molecule is coordinated by His$^{6.52}$ and seems to preclude a buried rotamer conformation of Met$^{3.36}$ in κ-OR that is observed in NOP, resulting in a deviation of more than 6 Å among the Cε atoms of the side chain in these two crystal structures (Fig. 4c). This Met$^{3.36}$ residue is conserved in all opioid receptors and makes extensive hydrophobic interactions with the corresponding ligands in both NOP and κ-OR. As a consequence, the 7-hydroxyisoquinoline head group of the κ-OR ligand JDTic is not able to penetrate as deeply into this area of the orthosteric pocket as compared with the benzofuran group of C-24. The 'reoriented' hydroxylated head group of JDTic is stabilized by a hydrogen bond interaction to a water molecule that is coordinated by the backbone carbonyl of Lys 227$^{5.39}$, potentially explaining the need for a tyrosine residue at the N terminus of dynorphin A. With modifications of Lys$^{5.39}$ to Ala$^{5.39}$ and His$^{6.52}$ to Gln$^{6.52}$ in the NOP receptor, remodelling of the binding pocket that includes a backbone shift in helix V, repacking of the Met$^{3.36}$ side chain and water rearrangements provides a likely explanation for selectivity in the message domain of the peptide ligands.

Perhaps most intriguing are the evolutionary differences between NOP and the other three classical opioid receptors (κ-OR, μ-OR and δ-OR). Despite high sequence identity between receptors, marked differences in ligand selectivity between these opioid receptors go in hand with substantial changes in the structure of their binding pockets. This situation is very different from other GPCR subfamilies (for example, β-adrenergic and muscarinic) in which different subtypes signal by the same ligands via highly conserved orthosteric pocket architectures. With structural data for κ-OR[5], μ-OR[6] and NOP now available, and the fourth (δ-OR) opioid receptor structure likely to come in the near future, one can begin to investigate the ligand structure activity relationships and evolutionary aspects of this receptor subfamily in greater detail.

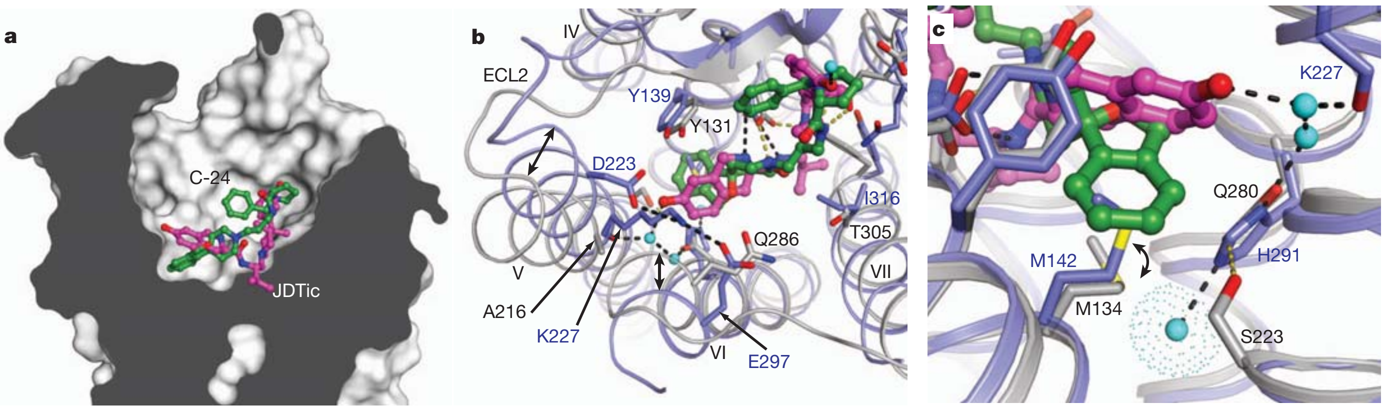


Figure 4 | Conformational differences in the ligand-binding pocket between NOP–C-24 and κ-OR–JDTic. a, 'Sliced' surface representation of NOP, highlighting the deep binding pocket bound with C-24 (coloured green) and JDTic (coloured magenta) from the superimposed κ-OR structure. **b**, **c**, Different views of NOP (coloured grey with green C-24) superimposed with the κ-OR structure[5] (PDB accession 4DJH; coloured blue with magenta JDTic). Hydrogen bonding interactions are depicted as dashed yellow and black lines for NOP and κ-OR, respectively. The water molecules from the κ-OR structure are depicted as cyan spheres. Residue labels are coloured black and blue for NOP and κ-OR, respectively. The conformational shifts observed between helices V and VI that result in different binding pocket architectures are highlighted in **b**. The alternative rotamer of Met$^{3.36}$ in the pocket (134 in NOP and 142 in κ-OR), which affects the orientation of the head group of the ligand, is highlighted in **c**.

METHODS SUMMARY

BRIL–NOP was expressed in *Spodoptera frugiperda* (Sf9) insect cells. Ligand-binding asays were performed as described in Methods. Sf9 membranes were solubilized using 0.5% (w/v) *n*-dodecyl-β-D-maltopyranoside and 0.1% (w/v) cholesteryl hemisuccinate, and purified by immobilized metal ion affinity chromatography. Receptor crystallization was performed by the lipidic cubic phase (LCP) method. The protein–LCP mixture contained 40% (w/w) concentrated receptor solution, 54% (w/w) monoolein and 6% (w/w) cholesterol. Crystals were grown in 40 nl protein-laden LCP bolus overlaid by 0.8 μl of precipitant solution (25–30% (v/v) PEG400, 100–200 mM potassium sodium tartrate tetrahydrate, 100 mM Bis-Tris propane, pH 6.4) at 20 °C. Crystals were collected directly from the LCP matrix and flash frozen in liquid nitrogen. X-ray diffraction data were collected at 100 K on the 23ID-B/D beamline (GM/CA-CAT) of the Advanced Photon Source at the Argonne National Laboratory using a 10-μm collimated minibeam. Diffraction data from 23 crystals were merged for the final data set. Data collection, processing, structure solution and refinement are described in Methods. Flexible docking of small molecules and peptides was performed with the ICM molecular modelling package (Molsoft LLC).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 22 February 2011; accepted 30 March 2012.

1. Mollereau, C. *et al.* ORL1, a novel member of the opioid receptor family. Cloning, functional expression and localization. *FEBS Lett.* **341,** 33–38 (1994).
2. Meunier, J. C. *et al.* Isolation and structure of the endogenous agonist of opioid receptor-like ORL1 receptor. *Nature* **377,** 532–535 (1995).
3. Reinscheid, R. K. *et al.* Orphanin FQ: a neuropeptide that activates an opioidlike G protein-coupled receptor. *Science* **270,** 792–794 (1995).
4. Goto, Y. *et al.* Identification of a novel spiropiperidine opioid receptor-like 1 antagonist class by a focused library approach featuring 3D-pharmacophore similarity. *J. Med. Chem.* **49,** 847–849 (2006).
5. Wu, H. *et al.* Structure of the human kappa opioid receptor in complex with JDTic. *Nature* advance online publication doi:10.1038/nature10939 (21 March 2012).
6. Manglik, A. *et al.* Crystal structure of the μ-opioid receptor bound to a morphinan antagonist. *Nature* advance online publication doi:10.1038/nature10954 (21 March 2012).
7. Lambert, D. G. The nociceptin/orphanin FQ receptor: a target with broad therapeutic potential. *Nature Rev. Drug Discov.* **7,** 694–710 (2008).
8. Chu, R. *et al.* Redesign of a four-helix bundle protein by phage display coupled with proteolysis and structural characterization by NMR and X-ray crystallography. *J. Mol. Biol.* **323,** 253–262 (2002).
9. Varty, G. B. *et al.* The anxiolytic-like effects of the novel, orally active nociceptin opioid receptor agonist 8-[bis(2-methylphenyl)methyl]-3-phenyl-8-azabicyclo[3.2.1]octan-3-ol (SCH 221510). *J. Pharmacol. Exp. Ther.* **326,** 672–682 (2008).
10. Fischetti, C. *et al.* Pharmacological characterization of the nociceptin/orphanin FQ receptor non peptide antagonist Compound 24. *Eur. J. Pharmacol.* **614,** 50–57 (2009).
11. Volta, M., Viaro, R., Trapella, C., Marti, M. & Morari, M. Dopamine-nociceptin/orphanin FQ interactions in the substantia nigra reticulata of hemiparkinsonian rats: involvement of D2/D3 receptors and impact on nigro-thalamic neurons and motor activity. *Exp. Neurol.* **228,** 126–137 (2011).
12. Chavkin, C. & Goldstein, A. Specific receptor for the opioid peptide dynorphin: structure–activity relationships. *Proc. Natl Acad. Sci. USA* **78,** 6543–6547 (1981).
13. Katritch, V., Cherezov, V. & Stevens, R. C. Diversity and modularity of G protein-coupled receptor structures. *Trends Pharmacol. Sci.* **33,** 17–27 (2012).
14. Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci.* **25,** 366–428 (1995).
15. Wu, B. *et al.* Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* **330,** 1066–1071 (2010).
16. Fredriksson, R., Lagerstrom, M. C., Lundin, L. G. & Schioth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **63,** 1256–1272 (2003).
17. Lapalu, S. *et al.* Comparison of the structure-activity relationships of nociceptin and dynorphin A using chimeric peptides. *FEBS Lett.* **417,** 333–336 (1997).
18. Doré, A. S. *et al.* Structure of the adenosine A_{2A} receptor in complex with ZM241385 and the xanthines XAC and caffeine. *Structure* **19,** 1283–1293 (2011).
19. Mouledous, L., Topham, C. M., Moisand, C., Mollereau, C. & Meunier, J. C. Functional inactivation of the nociceptin receptor by alanine substitution of glutamine 286 at the C terminus of transmembrane segment VI: evidence from a site-directed mutagenesis study of the ORL1 receptor transmembrane-binding domain. *Mol. Pharmacol.* **57,** 495–502 (2000).
20. Akuzawa, N., Takeda, S. & Ishiguro, M. Structural modelling and mutation analysis of a nociceptin receptor and its ligand complexes. *J. Biochem.* **141,** 907–916 (2007).
21. Topham, C. M., Mouledous, L., Poda, G., Maigret, B. & Meunier, J. C. Molecular modelling of the ORL1 receptor and its complex with nociceptin. *Protein Eng.* **11,** 1163–1179 (1998).
22. Trapella, C. *et al.* Structure-activity studies on the nociceptin/orphanin FQ receptor antagonist 1-benzyl-N-{3-[spiroisobenzofuran-1(3H),4'-piperidin-1-yl]propyl} pyrrolidine-2-carboxamide. *Bioorg. Med. Chem.* **17,** 5080–5095 (2009).
23. Guerrini, R. *et al.* Further studies on nociceptin-related peptides: discovery of a new chemical template with antagonist activity on the nociceptin receptor. *J. Med. Chem.* **43,** 2805–2813 (2000).
24. Calo, G. *et al.* [Nphe1,Arg14,Lys15]nociceptin-NH2, a novel potent and selective antagonist of the nociceptin/orphanin FQ receptor. *Br. J. Pharmacol.* **136,** 303–311 (2002).
25. Tancredi, T. *et al.* The interaction of highly helical structural mutants with the NOP receptor discloses the role of the address domain of nociceptin/orphanin FQ. *Chemistry* **11,** 2061–2070 (2005).
26. Orsini, M. J. *et al.* The nociceptin pharmacophore site for opioid receptor binding derived from the NMR structure and bioactivity relationships. *J. Biol. Chem.* **280,** 8134–8142 (2005).
27. Bes, B. & Meunier, J. C. Identification of a hexapeptide binding region in the nociceptin (ORL1) receptor by photo-affinity labelling with Ac-Arg-Bpa-Tyr-Arg-Trp-Arg-NH2. *Biochem. Biophys. Res. Commun.* **310,** 992–1001 (2003).
28. Meng, F. *et al.* Moving from the orphanin FQ receptor to an opioid receptor using four point mutations. *J. Biol. Chem.* **271,** 32016–32020 (1996).
29. Meng, F. *et al.* Creating a functional opioid alkaloid binding site in the orphanin FQ receptor through site-directed mutagenesis. *Mol. Pharmacol.* **53,** 772–777 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by PSI:Biology grant U54 GM094618 for biological studies and structure production, NIH Roadmap grant P50 GM073197 for technology development and R01 DA017204, R01 DA27170, and the NIMH Psychoactive Drug Screening Program (X.-P.H., E.V. and B.L.R.) and the Michael Hooker Chair of Pharmacology (B.L.R.), University of Ferrara (FAR grant to G.C.), Italian Ministry of University (FIRB Futuro in Ricerca 2010 grant to C.T.). We thank J. Francis for suggesting the idea to pursue the NOP receptor; J. Velasquez for help on molecular biology; T. Trinh, K. Allin and M. Chu for help on baculovirus expression; A. Walker and E. Abola for assistance with manuscript preparation; J. Smith, R. Fischetti and N. Sanishvili for assistance in development and use of the minibeam and beamtime at GM/CA-CAT beamline 23-ID at the Advanced Photon Source, which is supported by National Cancer Institute grant Y1-CO-1020 and National Institute of General Medical Sciences grant Y1-GM-1104.

Author Contributions A.A.T. optimized the constructs, purified and crystallized the receptor in LCP, optimized crystallization conditions, grew crystals for data collection, collected the data and refined the structure, and prepared the manuscript. W.L. assisted with LCP experiments, performed fluorescence recovery after photobleaching assays, collected diffraction data, and assisted with preparing the manuscript. E.C. assisted in construct optimization, assisted with LCP experiments, collected diffraction data, and assisted with preparing the manuscript. V.K. performed the receptor docking and prepared the manuscript. H.W. assisted with membrane preparations, provided advice on crystallization strategies, and assisted with preparing the manuscript. E.V. and X.-P.H. performed ligand-binding and site-directed mutagenesis studies. C.T., R.G. and G.C. suggested the use of and synthesized numerous ligands for crystallization and pharmacological studies, and assisted with preparing the manuscript. B.L.R. supervised the pharmacology and mutagenesis studies and prepared the manuscript. V.C. assisted with the crystallization in LCP, collected diffraction data, processed diffraction data, and prepared the manuscript. R.C.S. was responsible for the overall project strategy and management and wrote the manuscript.

Author Information The coordinates and the structure factors have been deposited in the Protein Data Bank under accession code 4EA3. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.C.S. (stevens@scripps.edu).

METHODS

Cloning, expression and purification. NOP contains a ~50-amino-acid extracellular domain at its N terminus, with a relatively high content of leucine and proline residues (26% and 14%, respectively) and three putative N-linked glycosylation sites. Despite high thermostability in the presence of select small molecule compounds, numerous attempts at crystallizing the receptor with an intact N terminus were unsuccessful. Although deletion of the C terminus (NOP-ΔC; 31-amino-acid deletion) resulted in increased expression, any truncation of the N terminus decreased the expression levels. However, replacement of the N terminus with several soluble fusion proteins restored the expression to levels that were comparable with constructs containing a full N terminus. Fusion with the thermostabilized apocytochrome b_{562}RIL (BRIL)[8] resulted in a construct (BRIL-ΔN-NOP-ΔC; referred to as BRIL–NOP in the manuscript) that was crystallized to 3.0 Å resolution in complex with the non-peptide antagonist C-24 (Banyu Pharmaceuticals).

The wild-type human NOP gene (encoded by *OPRL1*; UniProt accession P41146) was synthesized by DNA2.0 with codon optimization for *Spodoptera frugiperda* (*Sf9*), and then cloned into a modified pFastBac1 vector (Invitrogen) containing an expression cassette with a haemagglutinin signal sequence followed by a Flag tag at the N terminus, and a PreScission protease site followed by a 10×His tag at the C terminus. Thirty-one amino acids were deleted from the C terminus (residues 341–370), and 43 residues of the N terminus (residues 1–43) of NOP were replaced with the thermostabilized apocytochrome b_{562}RIL from *Escherichia coli* (M7W, H102I and K106L) (BRIL) protein using splicing by overlap extension PCR[30]. Recombinant baculoviruses were generated using the Bac-to-Bac system (Invitrogen) and were used to infect *Sf9* insect cells at a density of 2×10^6 cells ml^{-1} at a multiplicity of infection of 5. Infected cells were grown at 27 °C for 48 h before being collected, and the cell pellets were stored at −80 °C.

Insect cell membranes were disrupted by dounce homogenization of cell pellets in a hypotonic buffer containing 25 mM HEPES, pH 7.5, 10 mM $MgCl_2$, 20 mM KCl and protease inhibitor cocktail (Roche). Extensive washing of the membranes was performed consecutively by repeated dounce homogenization and centrifugation in the same hypotonic buffer (approximately once more), followed by high osmotic buffer containing 1.0 M NaCl, 10 mM $MgCl_2$, 20 mM KCl and 25 mM HEPES, pH 7.5 (three to four times). Purified membranes were resuspended in 500 mM NaCl, 20 mM KCl, 50 mM HEPES, pH 7.5, and 35% (v/v) glycerol, flash frozen with liquid nitrogen, and stored at −80 °C until further use.

Purified membranes were thawed and incubated with 25 μM C-24 (1-benzyl-N-{3-[spiroisobenzofuran-1(3H),4′-piperidin-1-yl]propyl} pyrrolidine-2-carboxamide)[4] (synthesized by C. Trapella and R. Guerrini), 500 mM NaCl, 20 mM KCl, 50 mM HEPES, pH 7.5, and 5% glycerol (v/v), and incubated at 4 °C for 1 h. Iodoacetamide (Sigma) was then added to the membranes at a final concentration of 1 mg ml^{-1} for another 15 min before solubilization with 0.5% (w/v) *n*-dodecyl-β-D-maltopyranoside (DDM; Anatrace), and 0.1% (w/v) cholesteryl hemisuccinate (CHS; Anatrace or Sigma) for 3 h at 4 °C. The supernatant was isolated by centrifugation at 160,000*g* for 45 min, supplemented with 25 mM imidazole, pH 7.5, and incubated with TALON metal ion affinity chromatography resin (Clontech) overnight at 4 °C. Typically, 0.75 ml of resin (slurry) per 1 l of original culture volume was used. After binding, the resin was washed with 15 column volumes of wash buffer 1 (500 mM NaCl, 20 mM KCl, 10 mM $MgCl_2$, 50 mM HEPES, pH 7.5, 5% (v/v) glycerol, 1 mM ATP, 25 mM imidazole, 25 μM C-24, 0.05% (w/v) DDM and 0.01% (w/v) CHS); and 5 column volumes of wash buffer 2 (same as wash buffer 1, but without ATP and $MgCl_2$), before protein elution with elution buffer (500 mM NaCl, 20 mM KCl, 50 mM HEPES, pH 7.5, 10% (v/v) glycerol, 250 mM imidazole, 25 μM C-24, 0.025% (w/v) DDM and 0.005% (w/v) CHS). Purified receptor was exchanged into a buffer containing 500 mM NaCl, 20 mM KCl, 5% (v/v) glycerol, 50 mM HEPES, pH 7.5, and 25 μM C-24 using a PD midiTrap G-25 column (GE Healthcare). BRIL-ΔN-NOP-ΔC was then supplemented with C-24 to a final concentration of 100 μM, and concentrated from ~0.4 mg ml^{-1} to 40 mg ml^{-1} with a 100-kDa molecular mass cut-off Vivaspin concentrator (GE Healthcare). Receptor purity and monodispersity was followed using SDS–PAGE and analytical size exclusion chromatography.

Pharmacological assays. The different NOP constructs (codon optimized for Sf9 expression) were cloned from pFastBac into pCDNA3.1 and expressed in HEK293T cells. Mutations (Q107A, D110A, D130A, Y131A, M134A, I219A, Q280A and Y309A) were introduced into the NOP sequence using standard QuikChange protocols. Binding affinity was determined from competition binding assays using ^{3}H-N/OFQ as a radioligand. NOP receptor-mediated inhibition of the cyclic AMP response was measured using a cAMP biosensor (see ref. 31 for details) to assess the functionality of the NOP constructs and the effects of alanine mutations within 3.5 Å from the antagonist C-24 in the structure. HEK293T cells were transiently transfected for binding assays or functional assays. Antagonist inhibition response curves were measured in the presence of a concentration of agonist (N/OFQ or SCH-221510) approximately corresponding to its EC_{80} value (the concentration that leads to an 80% maximum response). Results were analysed using GraphPad Prism.

Crystallization. Protein samples of BRIL-ΔN-NOP-ΔC (concentrated to 40 mg ml^{-1}) in complex with C-24 were reconstituted into the lipidic cubic phase (LCP) by mixing with molten lipid using a mechanical syringe mixer[32]. The protein–LCP mixture contained 40% (w/w) protein solution, 54% (w/w) monoolein (Sigma) and 6% (w/w) cholesterol (AvantiPolar Lipids). Crystallization trials were performed in 96-well glass sandwich plates[33] (Marienfeld) by the NT8-LCP (Formulatrix) or mosquito LCP (TTP LabTech) crystallization robots using 40 nl protein-laden LCP bolus overlaid with 0.8 μl precipitant solution in each well, and sealed with a glass coverslip. Protein reconstitution in LCP and crystallization trials were carried out at room temperature (~20–22 °C). The crystallization plates were stored and imaged in an incubator/imager (RockImager 1000, Formulatrix) at 20 °C. Diffraction quality crystals of an average size of 40 × 10 × 3 μm were obtained within ~14 days in 25–30% (v/v) PEG400, 100–200 mM potassium sodium tartrate tetrahydrate, 100 mM Bis-Tris propane, pH 6.4. Crystals were collected directly from LCP using 50 μm MiTeGen micromounts and immediately flash frozen in liquid nitrogen without adding extra cryoprotectant.

X-ray data collection and processing. Crystallographic data were collected on the 23ID-B/D beamline (GM/CA CAT) of the Advanced Photon Source at the Argonne National Laboratory, using a 10 μm collimated minibeam. Because of radiation damage, typically 10° of data was collected using an unattenuated beam, 1° oscillation and 3–10 s exposure before moving to a fresh part of the crystal, if possible, or changing the crystal. Partial data sets from 23 crystals were integrated, scaled and merged together using HKL2000 (ref. 34).

Structure determination and refinement. Initial molecular replacement solution was obtained by Phaser[35] using the receptor domain of the ΔN-κ-OR-T4L-ΔC/JDTic structure (PDB accession 4DJH)[5] and the thermostabilized apocytochrome b_{562}RIL protein (PDB accession 1M6T) as search models. With two antiparallel receptor molecules in the asymmetric unit of the $P2_1$ lattice, one of the BRIL domains is disordered whereas the second forms crystal lattice contacts with two receptors from an adjacent layer. We suspect that the disordered BRIL domain is flexible owing to the presence of a flexible linker and the absence of any crystal lattice contacts. The structure was refined by repetitive cycling between Coot[36] and Phenix[37]. The initial stages of refinement were performed with simulated annealing and rebuilding into composite omit maps, and noncrystallographic symmetry and translation/libration/screw (TLS) refinement were implemented throughout. The data collection and refinement statistics are shown in Supplementary Table 5. Figures were created using PyMOL[38], and electrostatic surface potentials were obtained using APBS[39].

Molecular modelling of C-24 analogues and UFP-101 peptide binding to NOP. Docking of high-affinity NOP specific ligands was performed using an all-atom flexible receptor docking algorithm in ICM-Pro (MolSoft LLC) molecular modelling package as described previously[40]. Internal coordinate (torsion) movements were allowed in the side chains of the binding pocket, defined as residues within 10 Å distance of C-24 in the crystal structure. Other side chains and the backbone of the protein were kept as in the crystal structure. An initial conformation for small molecule ligands was generated by Cartesian optimization of the ligand model in Merck Molecular Force Field. Docking was performed by placing the ligand in a random position within 5 Å from the entrance to the binding pocket and global conformational energy optimization of the complex[40,41]. To facilitate side-chain rotamer switches in flexible NOP receptor models, the first 10^6 steps of the Monte Carlo procedure used 'soft' van der Waals potentials and high Monte Carlo temperature, followed by another 10^6 steps with 'exact' van der Waals potentials and gradually decreasing temperature. A harmonic 'distance restraint' was applied between the protonated amine (of piperidine group in the small ligand or Nphe 1 in the UFP-101 peptide) and the carboxyl of the Asp $130^{3.32}$ side chain in the initial 10^6 steps. The restraint was removed in the final 10^6 steps of the docking procedure. With UFP-101, the first six residues Nphe 1-Gly 2-Gly 3-Phe 4-Thr 5-Gly 6 were considered fully flexible, whereas the peptide backbone was fixed in an ideal α-helical conformation for the rest of the peptide (Ala 7-Arg 8-Lys 9-Ser 10-Ala 11-Arg 12-Lys 13-Arg 14-Lys 15-Asn 16-Gln 17). At least 10 independent runs of the docking procedure were performed for each NOP-ligand. The docking results were considered 'consistent' when at least 80% of the individual runs resulted in conformations clustered within an r.m.s.d. of <1 Å to the overall best energy pose for small molecule ligands and within an r.m.s.d. of <2 Å for the UFP-101 peptide. All calculations were performed on a 12-core Linux workstation.

30. Heckman, K. L. & Pease, L. R. Gene splicing and mutagenesis by PCR-driven overlap extension. *Nature Protocols* **2,** 924–932 (2007).
31. Kimple, A. J. *et al.* Structural determinants of G-protein alpha subunit selectivity by regulator of G-protein signaling 2 (RGS2). *J. Biol. Chem.* **284,** 19402–19411 (2009).

32. Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nature Protocols* **4,** 706–731 (2009).
33. Cherezov, V., Peddi, A., Muthusubramaniam, L., Zheng, Y. F. & Caffrey, M. A robotic system for crystallizing membrane and soluble proteins in lipidic mesophases. *Acta Crystallogr. D* **60,** 1795–1807 (2004).
34. Otwinowski Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
35. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
36. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
37. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
38. The PyMOL Molecular Graphics System. v.1.4.1 (2011).
39. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA* **98,** 10037–10041 (2001).
40. Totrov, M. & Abagyan, R. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins* **29** (suppl.), 215–220 (1997).
41. Katritch, V. *et al.* Analysis of full and partial agonists binding to β_2-adrenergic receptor suggests a role of transmembrane helix V in agonist-specific conformational changes. *J. Mol. Recognit.* **22,** 307–318 (2009).

doi:10.1038/nature11111

Structure of the δ-opioid receptor bound to naltrindole

Sébastien Granier[1,2], Aashish Manglik[1]*, Andrew C. Kruse[1]*, Tong Sun Kobilka[1], Foon Sun Thian[1], William I. Weis[1,3] & Brian K. Kobilka[1]

The opioid receptor family comprises three members, the μ-, δ- and κ-opioid receptors, which respond to classical opioid alkaloids such as morphine and heroin as well as to endogenous peptide ligands like endorphins. They belong to the G-protein-coupled receptor (GPCR) superfamily, and are excellent therapeutic targets for pain control. The δ-opioid receptor (δ-OR) has a role in analgesia, as well as in other neurological functions that remain poorly understood[1]. The structures of the μ-OR and κ-OR have recently been solved[2,3]. Here we report the crystal structure of the mouse δ-OR, bound to the subtype-selective antagonist naltrindole. Together with the structures of the μ-OR and κ-OR, the δ-OR structure provides insights into conserved elements of opioid ligand recognition while also revealing structural features associated with ligand-subtype selectivity. The binding pocket of opioid receptors can be divided into two distinct regions. Whereas the lower part of this pocket is highly conserved among opioid receptors, the upper part contains divergent residues that confer subtype selectivity. This provides a structural explanation and validation for the 'message–address' model of opioid receptor pharmacology[4,5], in which distinct 'message' (efficacy) and 'address' (selectivity) determinants are contained within a single ligand. Comparison of the address region of the δ-OR with other GPCRs reveals that this structural organization may be a more general phenomenon, extending to other GPCR families as well.

Opioid receptors have an important role in the central nervous system, regulating pain perception, hedonic homeostasis, mood and wellbeing[1]. Thus, they have long been the focus of physiological and pharmacological studies, as well as being important therapeutic targets. The opioid receptors are GPCRs, and were classified based on their pharmacology and their tissue distribution into three subclasses: the μ (for morphine), the δ (for vas deferens) and the κ (for ketocyclazocine) receptors[6]. The sequence identity within the transmembrane domains (TMs) between the μ-OR and δ-OR, the μ-OR and κ-OR, and the δ-OR and κ-OR is 76%, 73% and 74%, respectively[7]. Another receptor identified by cloning, the nociceptin/orphanin FQ receptor, was classified in this family owing to a high sequence identity (67% in the TM)[7]. However, morphinans and most other classical opioid ligands have little affinity for the nociceptin receptor[8]. The μ-, δ- and κ-ORs are activated by endogenous peptides: the endorphins, enkephalins and dynorphins[8]. They are also the targets of chemically diverse small molecules with a variety of efficacy and selectivity profiles[8]. In an effort to understand better the structural basis for opioid receptor pharmacology and function, we used the *in meso* crystallization method to solve a 3.4 Å structure of a *Mus musculus* δ-OR T4 lysozyme (T4L) fusion protein (Supplementary Fig. 1) bound to naltrindole, a non-covalent δ-OR-selective morphinan antagonist[9].

The δ-OR structure presents the typical GPCR seven-pass transmembrane helix fold (Fig. 1a), and shows marked conservation of backbone structure with other opioid receptors (Fig. 1b, c), even in regions with low sequence conservation (Fig. 1d, e). The ligand naltrindole sits in an exposed binding pocket, similar in shape to that observed for the μ-OR and κ-OR[2,3]. The CXCR4 receptor also has a solvent-exposed binding pocket, suggesting that this may be a feature common to some GPCRs activated by peptides. The β-hairpin in extracellular loop (ECL)2 (Fig. 1d) is observed in all three opioid receptors, despite the low sequence identity in this domain. ECL3, which is also poorly conserved among the three opioid receptors, was unresolved in the κ-OR structure and has high temperature factors in both μ-OR and δ-OR (Fig. 1e), suggesting a relatively flexible link between TM6 and TM7. Of note, the κ-OR structure shows a clear difference in the position of the extracellular half of TM1, with an outward displacement of approximately 10 Å (Fig. 1b) compared to the μ-OR and δ-OR. In this respect, the μ-OR and δ-OR resemble each other and the CXCR4 chemokine receptor more closely than the κ-OR. However, this structural difference may simply reflect differences in crystallization conditions or crystal packing influences, as is seen in the turkey β_1-adrenergic receptor structure (PDB accession 2VT4), where two different TM1 conformations are observed in the same crystal[10].

δ- and μ-ORs have been observed to form homo-oligomers in transfected cells, and functional studies suggest that they form pharmacologically distinct hetero-oligomers when they are co-expressed[11]. It is therefore interesting that in the μ-OR crystal lattice two parallel dimeric interfaces were observed[2] (Supplementary Fig. 2). The most extensive interface involves TM5 and TM6 of adjacent protomers. The other interface, which is also observed in the κ-OR, involves TM1, TM2 and helix 8. In addition to this common interface, an anti-parallel interaction is also observed in the κ-OR crystal lattice. In contrast, the δ-OR crystallizes with only an anti-parallel arrangement of receptor molecules (Supplementary Fig. 2). However, inferences regarding the physiological relevance of oligomeric interfaces observed in GPCR crystal structures should be made with caution. Purified, detergent-solubilized δ-OR and μ-OR are monomeric before crystallization and the association into either parallel or antiparallel dimers occurs during crystallogenesis. The differences between the μ-OR and δ-OR dimeric interfaces probably reflect differences in the most energetically favourable interactions under the crystallography conditions and may be the consequence of one or more of the following: different crystallization conditions, a different T4L arrangement, sequence differences in the protein, and subtle differences in the structures stabilized by the different ligands.

Opioid receptors bind exceptionally well to highly diverse ligands, including morphinans, a wide variety of other small molecules, and peptides of varying length. Details of naltrindole binding to the δ-OR are presented in Fig. 2 and Supplementary Fig. 3. Despite their chemical diversity, many opioid ligands display conserved features, most notably a phenolic hydroxyl separated by six carbons from a positive charge, which mimics the amino-terminal tyrosine of all endogenous opioid peptides (Fig. 3). The morphinan ligand naltrindole used in

[1]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, California 94305, USA. [2]CNRS UMR 5203, and INSERM U661, and Université Montpellier 1 et 2, Institut de Génomique Fonctionnelle, Montpellier 34094, France. [3]Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305, USA.
*These authors contributed equally to this work.

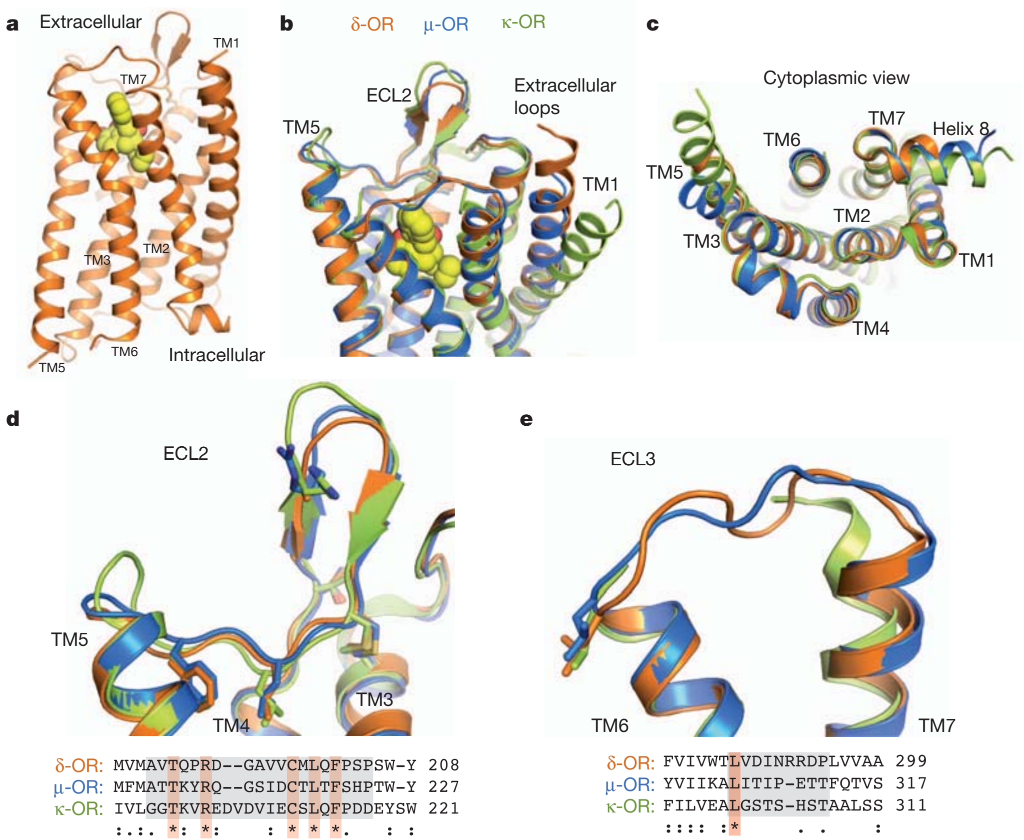


Figure 1 | Overall structure of the δ-OR. **a**, The δ-OR, orange, exhibits a typical seven-pass transmembrane architecture common to other GPCRs. **b**, **c**, This fold is highly conserved among all three classical members of the opioid receptor family. δ-OR, orange; μ-OR, blue; κ-OR, green. **d**, The opioid family possesses a conserved β-strand fold in ECL2, creating a wide, open binding pocket. Despite the structural similarity, only five residues in this region are absolutely conserved. Conserved residues are highlighted red in sequence alignment and shown as sticks. Asterisks indicate positions with complete residue conservation among opioid subtypes, colons indicate residues with strongly similar properties, and periods indicate residues with weakly similar properties. **e**, ECL3, an important selectivity determinant for ligand binding, shows modest structural variability in the μ-OR and δ-OR. In the κ-OR receptor structure this region could not be resolved owing to poor electron density. A single leucine residue is conserved in ECL3 among opioid subtypes.

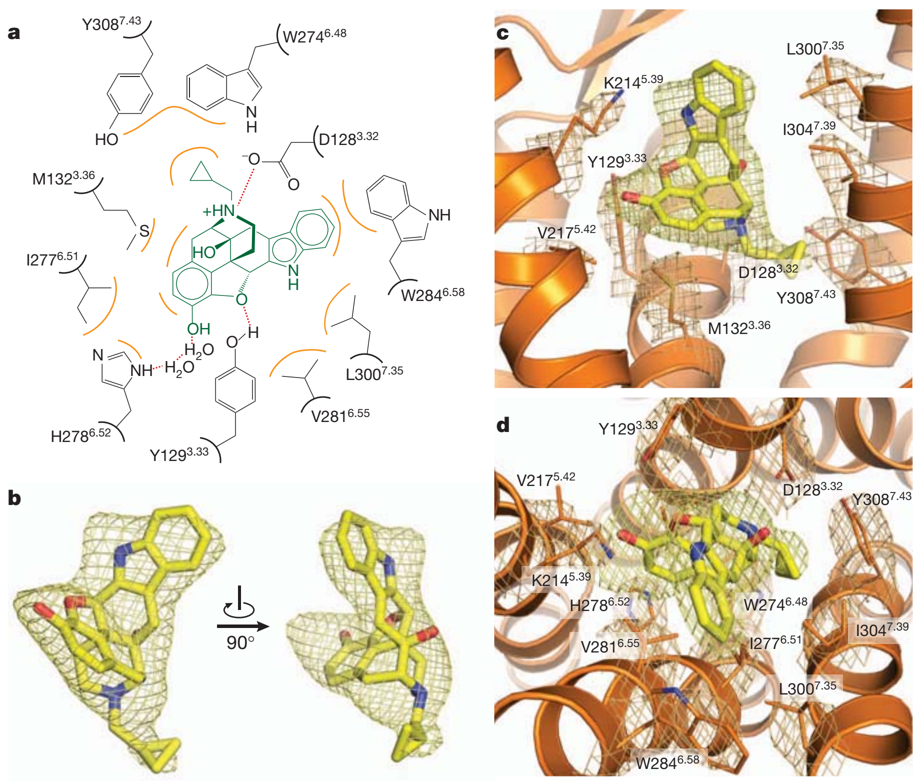


Figure 2 | Ligand-binding site of the δ-OR. **a–d**, Naltrindole binds in a deep but open binding site within the δ-OR. **a**, **b**, Selected contacts are highlighted (**a**), and a ligand $F_o - F_c$ omit map within a 2 Å radius of naltrindole is shown at a 3σ contour (**b**). **c**, **d**, The complete binding site is shown. $2F_o - F_c$ electron density maps within a 2 Å radius of binding site amino acid side chains are shown in orange at a 1.5σ contour. The ligand omit density is shown as in **b**.

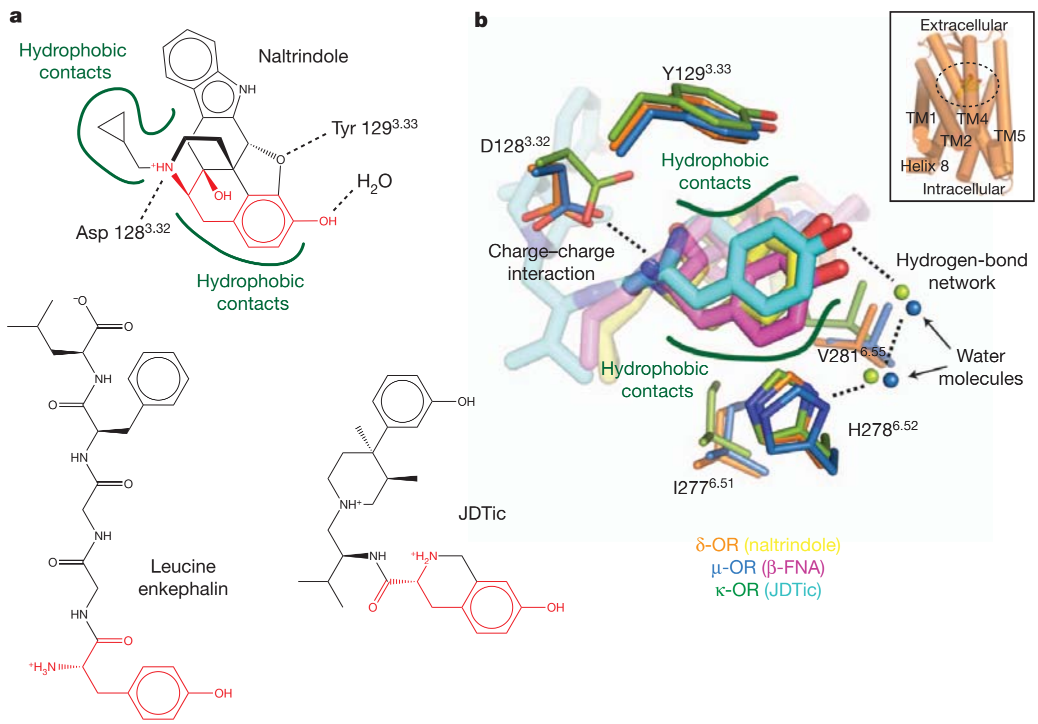


Figure 3 | A conserved opioid ligand recognition mode. **a**, Opioid receptors bind a wide variety of ligands, including morphinans like naltrindole, other small molecules such as JDTic, and peptides like enkephalins. Most opioid ligands, including these, contain a 'tyrosine' pharmacophore (red) with a phenolic hydroxyl in close proximity to a positive charge. Conserved recognition features for morphinan ligands observed in the μ-OR and δ-OR are highlighted on naltrindole (top). **b**, The tyrosine pharmacophores of naltrindole, β-FNA and JDTic are shown in their three-dimensional context as observed in the crystal structures of the δ-OR, μ-OR and κ-OR. The inset shows the receptor orientation in the expanded view. Specific conserved interactions in all three receptors are highlighted. The δ-OR is shown in orange (naltrindole in yellow), the μ-OR in dark blue (β-FNA in pink), and the κ-OR in green (JDTic in light blue).

crystallization of the δ-OR is non-covalent[9], and is therefore not subject to possible distortions in its binding mode due to a covalent tether like that in the structure of the μ-OR. Indeed, the position of naltrindole in the binding pocket is shifted slightly relative to the position of the covalent morphinan ligand β-funaltrexamine (β-FNA)[12] bound to the μ-OR (Fig. 4), although all major interactions are present in both structures. As anticipated from the μ-OR structure, the leucine in the position $300^{7.35}$ is in contact with the indole group of naltrindole (Fig. 2). This residue is responsible for naltrindole selectivity[13], as $W318^{7.35}$ in μ-OR and $Y312^{7.35}$ in κ-OR are sterically incompatible with naltrindole binding. As with the μ-OR and κ-OR, electron density near the phenolic hydroxyl of naltrindole suggests the presence of water molecules (Fig. 3b and Supplementary Fig. 4). The resolution of the δ-OR structure, however, is not sufficient to place solvent molecules with confidence. Nonetheless, this feature suggests that a two-water hydrogen-bond network linking $H278^{6.52}$ and the ligand phenolic hydroxyl is probably a conserved feature of opioid ligand recognition, and certainly this appears to be the case in the μ-OR and κ-OR.

With the structure of the δ-OR, all classical opioid receptors have now been crystallized and solved in inactive conformations. A closely related receptor, that for the nociceptin/orphanin FQ peptide, is often classified within the opioid receptor family. However, this receptor has low or negligible affinity for most opioid alkaloids[8], despite high sequence conservation within the TM domains. Examination of the morphinan binding site of δ-OR reveals that only a few of the critical interacting residues differ between δ-OR and the nociceptin receptor (Supplementary Fig. 5). However, mutation of certain residues in the nociceptin receptor to their δ-OR counterparts can create a high-affinity alkaloid-binding site[14]. These mutations change smaller amino acid side chains in the nociceptin receptor to larger residues in the corresponding positions of the δ-OR, so it is likely that the binding pocket of the nociception receptor is somewhat enlarged relative to that of the δ-OR. The loss of tightly fitting hydrophobic interactions with the morphinan ring of opioid alkaloids may therefore account, at least in part, for the much higher affinity of most morphinans for the δ-OR, μ-OR and κ-OR than for the nociceptin receptor.

Opioid pharmacology has long been described in terms of the 'message–address' concept[4,5], in which the ligand can be viewed as composed of two distinct modules carrying information about efficacy (message) and selectivity (address). The structure of the δ-OR and other opioid receptors reveals this pharmacological phenomenon to be a direct consequence of opioid receptor structure. The lower portion of the binding pocket is well-conserved in both sequence and structure. In the δ-OR, this portion of the binding pocket recognizes the core morphinan group, which entails the 'message' of the ligand (Fig. 4a, b). In contrast, the upper binding pocket is divergent among subtypes, and is rich in selectivity determinants (Fig. 4a, b). The indole 'address' of naltrindole extends into this region, conferring its δ-OR selectivity (Fig. 4a, b). Similarly, the 3-hydroxyphenyl group of JDTic extends into the address region, and would clash with δ-OR residue $K108^{2.63}$. This feature may account for the selectivity of JDTic for the κ-OR and μ-OR (Fig. 4a), although it is likely that other factors contribute as well.

Development of highly subtype-selective ligands has proven to be possible for the classical opioid receptors[15]. However, for another GPCR family, the muscarinic acetylcholine receptors, this has proven considerably more challenging. Comparison of the message and address regions of the δ-OR with the M2 muscarinic receptor[16] (Fig. 4c, d) reveals that the address region corresponds to the allosteric site of these receptors. This region is separated by a layer of tyrosines from the highly conserved orthosteric site, which matches the message region of opioid receptors. The physical separation of the two regions may therefore explain the relatively greater challenges associated with development of selective muscarinic ligands compared to opioid receptors, as well as the promising results of selective 'dual-steric' or

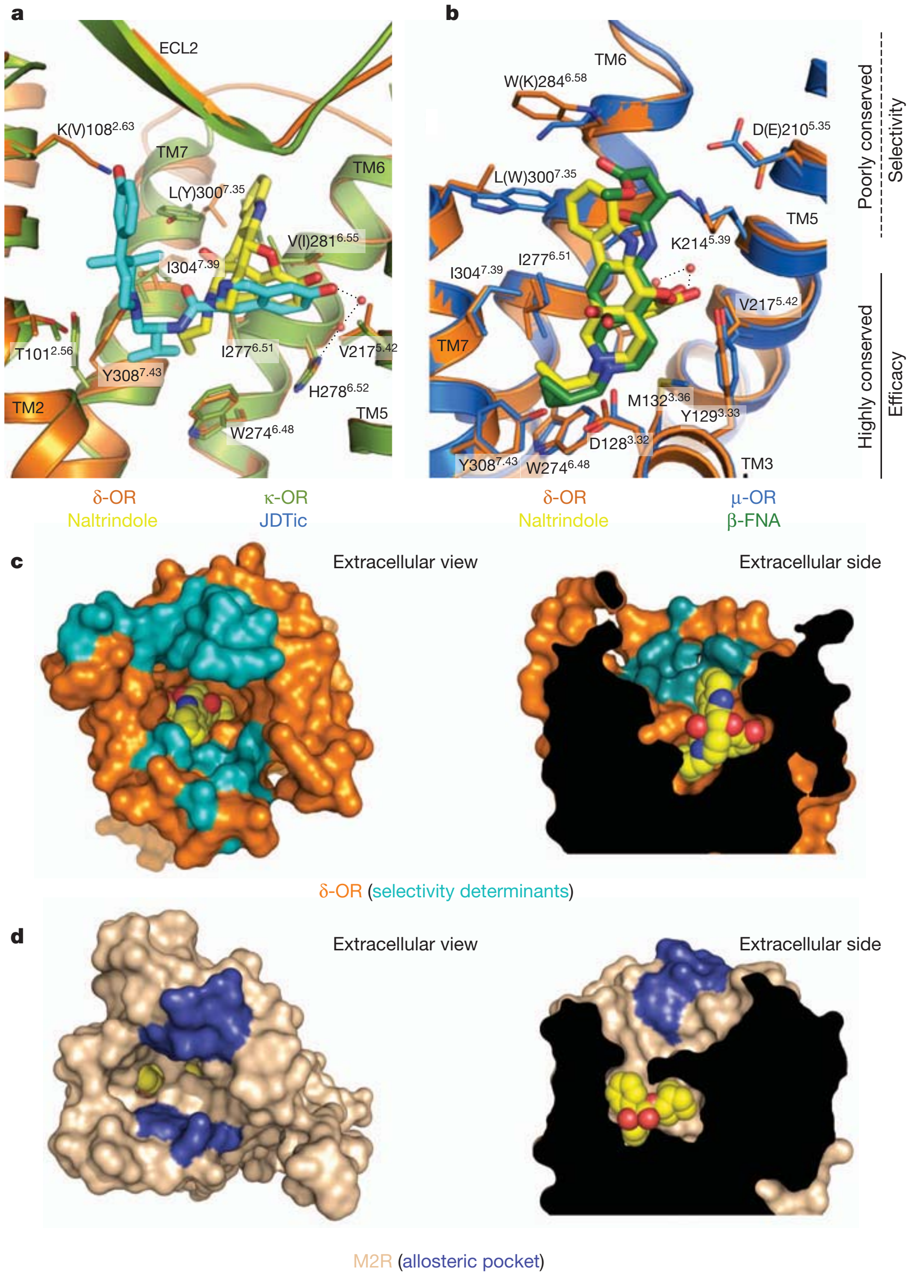


Figure 4 | The message–address hypothesis is reflected in opioid receptor structure. a, **b**, Comparison of the δ-OR with the κ-OR (**a**) and with the μ-OR (**b**) reveals high conservation in both sequence and structure in the base of the ligand-binding pocket, whereas extracellular regions are more divergent. These regions are delineated in the legend on right. The δ-OR is shown in orange (naltrindole in yellow), the κ-OR in light green (JDTic in light blue), and the μ-OR in dark blue (β-FNA is dark green). Residue numbers and labels are those for the δ-OR, with κ-OR residues in parentheses in **a** and μ-OR residues in parentheses in **b**, where sequence differs from the δ-OR. These regions interact with ligand moieties that can target binding to a particular opioid subtype. **c**, Residues previously characterized as important for opioid-subtype selectivity[18–20] are clustered around the upper part of the binding pocket, delineating an 'address' region of the receptor. The δ-OR is shown in orange (selectivity determinants in teal). **d**, This region is structurally analogous to the allosteric site in muscarinic receptors (allosteric pocket in blue), suggesting that the high selectivity of muscarinic ligands occupying this space is also a manifestation of the message–address features structurally encoded within GPCRs.

'bitopic' ligands that occupy both orthosteric and allosteric sites simultaneously[17]. The distinct message and address regions of the δ-OR and other opioid receptors then seem to be a more general feature of GPCRs, which may have implications for the development of ligands even for distantly related GPCR families.

Together with the μ-OR and κ-OR, the structure of the δ-OR completes the initial structural characterization of the opioid receptors, offering the first experimental views of the atomic details of ligand recognition and selectivity in this important GPCR family. However, such antagonist-bound structures are only the first step towards a complete understanding of opioid receptor function. Given the importance of opioid agonists in clinical medicine, active state structures, as well as signalling complexes, will be required to fully leverage structural methods towards the development of a new generation of opioid drugs.

METHODS SUMMARY

The δ-OR–T4L fusion protein was expressed in Sf9 insect cells and purified by nickel affinity chromatography followed by Flag antibody affinity chromatography and size exclusion chromatography. It was crystallized using the lipidic mesophase technique, and diffraction data were collected at GM/CA-CAT beamline 23ID-B at the Advanced Photon Source at Argonne National Laboratory. The structure was solved by molecular replacement using merged data from 20 crystals.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 27 March 2012; accepted 11 April 2012.

1. Pradhan, A. A., Befort, K., Nozaki, C., Gaveriaux-Ruff, C. & Kieffer, B. L. The delta opioid receptor: an evolving target for the treatment of brain disorders. *Trends Pharmacol. Sci.* **32,** 581–590 (2011).
2. Manglik, A. *et al.* Crystal structure of the μ-opioid receptor bound to a morphinan antagonist. *Nature* http://dx.doi.org/10.1038/nature10954 (this issue).
3. Wu, H. *et al.* Structure of the human κ-opioid receptor in complex with JDTic. *Nature* http://dx.doi.org/10.1038/10939 (this issue).
4. Chavkin, C. & Goldstein, A. Specific receptor for the opioid peptide dynorphin: structure–activity relationships. *Proc. Natl Acad. Sci. USA* **78,** 6543–6547 (1981).
5. Lipkowski, A. W., Tam, S. W. & Portoghese, P. S. Peptides as receptor selectivity modulators of opiate pharmacophores. *J. Med. Chem.* **29,** 1222–1225 (1986).
6. Satoh, M. & Minami, M. Molecular pharmacology of the opioid receptors. *Pharmacol. Ther.* **68,** 343–364 (1995).
7. Mollereau, C. *et al.* ORL1, a novel member of the opioid receptor family. Cloning, functional expression and localization. *FEBS Lett.* **341,** 33–38 (1994).
8. Cox, B. M. *et al. Opioid Receptors: Introduction* http://www.iuphar-db.org/DATABASE/FamilyIntroductionForward?familyId=50 (2009).
9. Portoghese, P. S., Sultana, M., Nagase, H. & Takemori, A. E. Application of the message-address concept in the design of highly potent and selective non-peptide δ opioid receptor antagonists. *J. Med. Chem.* **31,** 281–282 (1988).
10. Warne, T. *et al.* Structure of a β_1-adrenergic G-protein-coupled receptor. *Nature* **454,** 486–491 (2008).
11. George, S. R. *et al.* Oligomerization of μ- and δ-opioid receptors. Generation of novel functional properties. *J. Biol. Chem.* **275,** 26128–26135 (2000).
12. Portoghese, P. S., Larson, D. L., Sayre, L. M., Fries, D. S. & Takemori, A. E. A novel opioid receptor site directed alkylating agent with irreversible narcotic antagonistic and reversible agonistic activities. *J. Med. Chem.* **23,** 233–234 (1980).
13. Bonner, G., Meng, F. & Akil, H. Selectivity of μ-opioid receptor determined by interfacial residues near third extracellular loop. *Eur. J. Pharmacol.* **403,** 37–44 (2000).
14. Meng, F. *et al.* Creating a functional opioid alkaloid binding site in the orphanin FQ receptor through site-directed mutagenesis. *Mol. Pharmacol.* **53,** 772–777 (1998).
15. Eguchi, M. Recent advances in selective opioid receptor agonists and antagonists. *Med. Res. Rev.* **24,** 182–212 (2004).
16. Haga, K. *et al.* Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature* **482,** 547–551 (2011).
17. Valant, C. *et al.* A novel mechanism of G protein-coupled receptor functional selectivity. Muscarinic partial agonist McN-A-343 as a bitopic orthosteric/allosteric ligand. *J. Biol. Chem.* **283,** 29312–29321 (2008).
18. Metzger, T. G., Paterlini, M. G., Ferguson, D. M. & Portoghese, P. S. Investigation of the selectivity of oxymorphone- and naltrexone-derived ligands via site-directed mutagenesis of opioid receptors: exploring the ''address'' recognition locus. *J. Med. Chem.* **44,** 857–862 (2001).
19. Xue, J. C. *et al.* Differential binding domains of peptide and non-peptide ligands in the cloned rat κ opioid receptor. *J. Biol. Chem.* **269,** 30195–30199 (1994).
20. Pepin, M. C., Yue, S. Y., Roberts, E., Wahlestedt, C. & Walker, P. Novel ''restoration of function'' mutagenesis strategy to identify amino acids of the δ-opioid receptor involved in ligand binding. *J. Biol. Chem.* **272,** 9260–9267 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Sunahara for helpful suggestions on opioid receptor pharmacology and biochemistry. We acknowledge support from INSERM (S.G.), the Stanford Medical Scientist Training Program (A.M.), the National Science Foundation (A.C.K.), National Institutes of Health grants NS028471 (B.K.K.) and DA031418 (B.K.K.), and from the Mathers Foundation (B.K.K. and W.I.W.).

Author Contributions A.M., A.C.K. and S.G. designed experiments, performed research and analysed data. T.S.K. and F.S.T. expressed and purified receptor. W.I.W. supervised diffraction data analysis and model refinement. A.M., A.C.K., S.G. and B.K.K. prepared the manuscript. S.G. and B.K.K. supervised the research.

Author Information Coordinates and structure factors for δ-OR–T4L are deposited in the Protein Data Bank under accession code 4EJ4. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.G. (granier@stanford.edu) or B.K.K. (kobilka@stanford.edu).

METHODS

Expression and purification. We generated a *Mus musculus* δ-OR construct with features designed to enhance crystallogenesis. A tobacco etch virus (TEV) protease recognition site was introduced after residue 35, and the carboxy terminus was truncated after P342. T4L residues 2–161 were inserted in the third intracellular loop of δ-OR between residues 244 and 251. A Flag epitope tag was added to the N terminus and an octa-histidine tag was appended to the C terminus. The mouse and human δ-OR share 94% sequence identity, with most sequence differences in the disordered N and C termini. The final crystallization construct (δ-OR–T4L) is shown in Supplementary Fig. 1.

The δ-OR–T4L construct was expressed in Sf9 cells using the pFastBac (Invitrogen) baculovirus system in the presence of 10 μM naloxone. Cell cultures were grown to a density of 4×10^6 cells per ml, infected with baculovirus containing the δ-OR–T4L gene, shaken at 27 °C for 48 h, and cell pellets were harvested and stored at −80 °C. To purify the protein, insect cells were first lysed by osmotic shock in a buffer comprised of 10 mM Tris-HCl pH 7.5, 1 mM EDTA, 1 μM naltrindole and 2 mg ml^{-1} iodoacetamide to block reactive cysteines. This was followed by an extraction step, in which Sf9 membranes were homogenized with a glass dounce homogenizer in a solubilization buffer comprised of 1.0% lauryl maltose neopentyl glycol (MNG), 0.3% sodium cholate, 0.03% cholesterol hemisuccinate (CHS), 20 mM HEPES pH 7.5, 0.5 M NaCl, 30% v/v glycerol, 2 mg ml^{-1} iodoacetamide, and 1 μM naltrindole. This extraction reaction was mixed at 4 °C for 1 h, then centrifuged at high speed to remove cell debris. Nickel-NTA agarose was then added to the supernatant and stirred for 2 h. The Nickel-NTA resin was washed three times in batch with a washing buffer of 0.1% MNG, 0.03% sodium cholate, 0.01% CHS, 20 mM HEPES pH 7.5, 0.5 M NaCl and 1 μM naltrindole. The resin was transferred into a wide-bore glass column and bound receptor was eluted in washing buffer supplemented with 300 mM imidazole. Ni-NTA-purified δ-OR–T4L was then loaded over anti-Flag M1 affinity resin and the salt concentration was gradually lowered from 0.5 M to 0.1 M in a buffer otherwise comprised of 0.1% MNG, 0.01% CHS, 20 mM HEPES pH 7.5 and 1 μM naltrindole. The receptor was then washed with a buffer containing 0.01% MNG, 0.001% CHS, 20 mM HEPES pH 7.5, 0.1 M NaCl and 1 μM naltrindole and eluted from the anti-Flag M1 affinity resin with the same buffer containing 0.2 mg ml^{-1} Flag peptide and 2 mM EDTA. To remove flexible N and C termini, TEV protease was added at a 1:3 TEV:δ-OR–T4L ratio by weight. The sample was incubated at room temperature (23 °C) for 1 h followed by treatment with carboxypeptidase A (1:100 w/w) at 4 °C overnight. We used size exclusion chromatography to remove TEV and carboxypeptidase A. Size exclusion chromatography was performed on a Sephadex S200 column (GE Healthcare) in a buffer of 0.01% MNG, 0.001% CHS, 100 mM NaCl, 20 mM HEPES pH 7.5 and 1 μM naltrindole. After size exclusion, naltrindole was added to a final concentration of 10 μM. The resulting receptor preparation was pure and monodisperse (Supplementary Fig. 6).

Crystallization and data collection. Purified δ-OR–T4L receptor was concentrated to 50 mg ml^{-1} using a Vivaspin sample concentrator with a 50 kDa molecular weight cut-off (GE Healthcare). As for other GPCR–T4L fusion proteins crystallized so far, we used the *in meso* method to obtain crystals of δ-OR–T4L. Briefly, δ-OR–T4L was reconstituted into a mixture of monoolein and cholesterol (Sigma) by the two-syringe method. By weight, one part δ-OR–T4L was mixed with 1.5 parts of a 10:1 mixture of monoolein:cholesterol until the resulting phase was optically transparent. We used a Gryphon LCP robot (Art Robbins Instruments) to accurately dispense 20–55 nl mesophase drops onto glass plates. These lipidic boluses were overlaid with 700 nl precipitant solution. Crystals grew in precipitant solution consisting of 29–33% PEG 400, 100 mM HEPES pH 7.5, 120–180 mM sodium citrate (tribasic) and 350 mM magnesium chloride. Crystals were observed after 2 h and grew to full size after 5 days. Crystals used for data collection are shown in Supplementary Fig. 7.

Diffraction data were collected at Advanced Photon Source GM/CA-CAT beamline 23ID-B using a beam size of 10 μm. Owing to radiation damage, the diffraction quality decayed during exposure. Wedges of 5–15 degrees were collected and merged from 20 crystals using HKL2000[21]. Diffraction quality ranged from 3.0–3.5 Å in most cases. Due to anisotropic diffraction (see Supplementary Table 1) the highest shell $\langle I \rangle / \langle \sigma I \rangle$ value was slightly lower than is typical for isotropically diffracting crystals.

The structure of the δ-OR was solved by molecular replacement in Phaser[22] using the μ-OR receptor as a search model. The lattice for δ-OR–T4L shows alternating lipidic and aqueous layers, with receptor molecules arranged in anti-parallel associations (Supplementary Fig. 8). We improved the initial model by iteratively building regions of the receptor in Coot[23] and refining in Phenix[24]. To assess the quality of the final structure, we used MolProbity[25]. As with the μ-OR and κ-OR, electron density was clear, and allowed confident placement of the ligand and binding site residues (Supplementary Fig. 9). The resulting statistics for data collection and refinement are shown in Supplementary Table 1. Figures were prepared in PyMOL[26].

21. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
22. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Cryst.* **40,** 658–674 (2007).
23. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
24. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
25. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
26. The. PyMOL Molecular Graphics System v. 1.5.0.1 (Schrödinger, LLC, 2012).